DOCUMENT RESUME

ED 449 173 TM 032 253

AUTHOR Webber, Bev, Comp.
TITLE Examining Assessment. Conference Proceedings (Wellington, New Zealand, October 1999).
INSTITUTION New Zealand Council for Educational Research, Wellington.
ISBN ISBN-1-877140-76-7
PUB DATE 2000-00-00
NOTE 99p.
AVAILABLE FROM New Zealand Council for Educational Research Distribution Services, P.O. Box 3237, Wellington, New Zealand.
PUB TYPE Collected Works - Proceedings (021)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Educational Assessment; Educational Research; Elementary Secondary Education; Foreign Countries; Indigenous Populations; Maori (People); *Student Evaluation; *Test Use
IDENTIFIERS *New Zealand

ABSTRACT

Papers presented at the annual meeting of the New Zealand Council for Educational Research present diverse views on assessment in the context of proposals for national testing in primary schools in New Zealand. Also of widespread interest as the conference was developed was the restructuring of secondary school examinations. The papers are: (1) "Pressures, Priority and Progress: An International Perspective on School Assessment: Back to Basics" (Patrick Griffin); (2) "Commentary on Patrick Griffin's Paper" (Cedric Croft); (3) "A Primary Perspective" (Frances Nelson); (4) "Assessment Priorities for the Secondary Sector" (Graeme Macann); (5) "Assessment--Maori Attitudes to Assessment" (Wharehuia Hemara); (6) "Ministry of Education's Policy Interests in Assessment" (Frances Kelly); (7) "Achievement 2001" (Michael Irwin); and (8) "Conclusion" (Richard Watkins). Most of the papers contain references. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

EXAMINING ASSESSMENT

ED 449 173

Conference Proceedings

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Bev Webber

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM032253

NEW ZEALAND COUNCIL FOR EDUCATIONAL RESEARCH
TE RŪNANGA O AOTEAROA MŌ TE RANGAHAU I TE MĀTAURANGA

BEST COPY AVAILABLE

OCTOBER 1999

EXAMINING ASSESSMENT

CONFERENCE PROCEEDINGS
OCTOBER 1999

Compiled by Bev Webber

New Zealand Council for Educational Research
Wellington
2000

New Zealand Council for Educational Research
PO Box 3237
Wellington

© 2000 Copyright is held by individual writers over their own work.
Publishing rights are held by NZCER.

ISBN 1–877140–76–7

Distributed by NZCER Distribution Services
PO Box 3237
Wellington
New Zealand

CONTENTS

	Page
Foreword *Richard Watkins*	1
Pressures, Priority and Progress: An International Perspective on School Assessment: Back to Basics *Patrick Griffin*	3
Commentary on Patrick Griffin's Paper *Cedric Croft*	41
A Primary Perspective *Frances Nelson*	47
Assessment Priorities for the Secondary Sector *Graeme Macann*	53
Assessment—Maori Attitudes to Assessment *Wharehuia Hemara*	59
Ministry of Education's Policy Interests in Assessment *Frances Kelly*	67
Achievement 2001 *Michael Irwin*	77
Conclusion *Richard Watkins*	105

Note: *These papers have been published as presented. There has been no substantial editing, but where possible a standard layout has been used. The papers present a diversity of views on assessment.*

FOREWORD

Early in 1999, when NZCER staff first considered the theme of the forthcoming annual conference, a number of topics were discussed. Among these, assessment and reporting loomed large. By the time NZCER staff began to plan the 1999 conference, proposals to trial national testing in some primary schools and issues to do with restructuring secondary school examinations were topics of widespread interest. Accordingly, Cedric Croft from NZCER led the development of a programme and the selection of keynote speakers who could discuss aspects of these major themes.

In selecting the keynote speakers, NZCER was mindful that it needed to present a diversity of views, to illuminate assessment and to inspire further debate among teachers and the public about the application, utility and worth of assessment as a set of diagnostic and summative arrows in the educational quiver.

The ideas surrounding assessment, and to an extent measurement and evaluation, permeate much of the discussion in education. And it seems that issues associated with assessment in schools, particularly frequency, quality and utility, are never far from the attention of the media. So these issues are also the very stuff of contemporary political comment and observation. They are often wrapped in supposition, misunderstanding and polemic. Even amid the most quiescent group, mention assessment and the concomitant issue of standards, and a lively debate, though not necessarily an informed one, will almost inevitably occur.

Attention, not just in New Zealand, but in Australia, Britain, the USA and elsewhere, is increasingly given to improving the quality of assessment. This is critical, for it is only through quality assessment that students, teachers and the public at large can be assured that what has been taught has been satisfactorily learned, and that students are prepared for the next stages in their education. Most of the effort to improve assessment quality is driven by attention to the systematic specification of achievement levels in performance assessment.

It is important to state, however, that the term *assessment* is polysemic, having many meanings or applications. Thus it should be viewed according to context and application. And while it is polysemic, the term assessment can and should be quite clearly distinguished from measurement and evaluation, which perform quite different functions, and which carry with them their own contexts and applications.

Assessment may have formative or summative uses, but it is fair to say that in many western democracies it is the summative aspect of assessment which has received the most attention, certainly insofar as the media and many politicians are concerned. It is also fair to say that this emphasis has often occurred at the expense of developments which might have taken place in ensuring good formative assessment in a nation's schools, developments which would embrace diagnostic assessment by teachers of students, and by the students themselves.

Indeed, NZCER's long-established research and development work on a range of standardised tests, and more recently in the Assessment Resource Banks, has aimed to give teachers quality resources that may help underpin sound formative assessment.

It is also important to remember that assessment may comprise a wide array of tasks, from simple multiple choice "tick the box" tests, to complex and multiple judgments about the intrinsic worth or merit of an art work, performance or sporting contest.

Good assessment can take many forms, just as poor assessment can. Many classroom teachers may prove to be nearly as capable, and ultimately as capable, with appropriate training, of designing assessment instruments as are large corporations. Price is not always an issue: the fact that something costs a great deal is no guarantee that it will perform well, or that it will be useful in all circumstances.

No one method of assessment is inherently superior to any other. Each has its strengths and weaknesses. It is only through effective combinations of tasks, matched to clear purposes, that we can be assured that the result is a quality assessment upon which we may rely. Thus, no one approach should be dismissed because of inherent prejudices, or embraced because it has been popularised.

Effective assessment in New Zealand's schools and institutions must deploy an array of methods and approaches, depending on what is required, by whom, and for what purpose. If this occurs, then it seems inevitable that standards of teaching and learning will rise, along with public confidence in educational outcomes.

Richard Watkins
Director and Chief Executive
New Zealand Council for Educational Research
Wellington

PRESSURES, PRIORITY AND PROGRESS: AN INTERNATIONAL PERSPECTIVE ON SCHOOL ASSESSMENT: BACK TO BASICS

Patrick Griffin

Assessment Research Centre
Faculty of Education
The University of Melbourne

Reforming accountability in Australian education has been characterised by experimentation, radical thinking at a system level and changes that have kept its education systems at the forefront of world developments in education reform movements. Constitutional state control of education has perhaps hindered national developments, but this may be the very thing that gives states such as Victoria the freedom to take risks and chart new directions. The current development of self-managed schools, with the accountability that accompanies such a direction, had its roots in reforms that were introduced in the 1970s. When school and curriculum reform were introduced at that time, schools were encouraged though federal and state programs to innovate and find new and exciting ways to make school a better place and learning more effective. Curriculum support, relaxation of central examinations, the role of inspectors, and an emphasis of process over outcomes, led to an eventual loss of contact by schools with system administrators and the community.

From the mid 1970s to the mid 1980s, a succession of national and international studies presented results that suggested that learning was not being enhanced by the emphasis on process with scant attention to outcomes. Increasing retention rates and changing employment patterns for school leavers reinforced a growing conviction in the community that "standards" were falling. Victoria's change of state government in 1983 presented an opportunity to reassert system control over curriculum and to seek ways of monitoring achievements. Coincident with this were calls from business for more accountability in schools and for monitoring programs to be introduced. Teachers had enjoyed almost 20 years of freedom to teach and control the curriculum without any form of accountability, and this autonomy had become a "professional right" that would not be surrendered easily. Attempts to introduce national sample tests in 1982, for example, were abandoned after fierce campaigns by teacher unions.

The strategy employed was to introduce curriculum advisory documents (Curriculum Frameworks), backed by a series of Ministerial Papers (Ministerial Paper 6), and to include in them guidelines for assessment and evaluation. During the time of autonomy, teachers discovered a range of assessment methods immediately applicable to their classroom and abandoned almost all use of systemic tests. Despite the general impression that tests were inappropriate, sales of tests from ACER during this period were maintained at levels that suggested that teachers needed and used tests.

Curriculum frameworks brought control of the school curriculum back to the central system under an umbrella of rhetoric emphasising devolution of responsibility. At the same time, under Ministerial direction, the State Board of Education published a series of papers and conducted a series of consultancy seminars on assessment and accountability. These papers contained the seeds of an idea for test-free accountability. Encouraged by the State Board, the Victorian Ministry of Education developed the Literacy and Numeracy Profiles. These were immediately put to use in projects for monitoring the literacy and numeracy gains of federally funded programs.

In 1988 the Victorian Literacy Profiles were presented to the Australian Education Council as a model for national monitoring procedures. Eventually the model was adopted, and a national consortium of states and territories developed profiles covering an agreed series of learning areas. These profiles became statements of learning outcomes to be achieved by all students and were aligned to specific grade levels. Almost

immediately, the States began planning to monitor against the profile levels using tests. In an ironic twist, the test-free monitoring procedures, based on profiles, became the basis for standardised testing programs in all states and territories. In NSW, basic skills tests were introduced amid serious protests. All states eventually implemented testing programs. By 1995, it was clear that the interpretations of the national profiles differed across states and that there was no parity of achievement across systems. The federal government began to promote benchmarking of schools and students against baseline levels of achievement. The benchmark standards were developed between 1996 and 1999 and soon became the basis for state comparisons of achievement.

Australia does not have a single education system. Its constitution makes education the responsibility of the seven states and territories. Nevertheless the need for a national approach has been evident for many years. In the late 1980s, the Australian education systems agreed to define a common set of learning outcomes after the Australian Education Council examined ways of monitoring attainment on a national basis. The new learning outcomes were called Profiles. Their adoption was not a simple or easy process. Eight learning areas were defined by the Curriculum Advisory Committee (CURASS) and the process of national profiles began. The areas were English, science, mathematics, languages other than English (LOTE), technology, studies of society and environment, the arts, health (physical education and personal development). Additional unsuccessful pressure was brought to bear to have business education established as a learning area. Special areas of development included English as a second language, and special education. Cross curricular areas included the environment, information technology, personal and interpersonal skills, career and work education, literacy and numeracy. Groups with special needs included girls, Aboriginal and Torres Strait Islanders, isolated groups, children in poverty and early school leavers. Despite the good will intended to incorporate these issues in the statements and profiles, it proved difficult for the writers to develop threads of cross curricular themes in a curriculum based in the tightly structured discipline classifications.

Marsh (1995) described the reluctance of state systems to surrender local curriculum guidelines for the benefits of a national system for reporting student achievements:

> Whether the early checks and caveats about states' rights were ignored during the extensive activities to produce national profiles to be used for reporting within and across states, is a matter of conjecture (1995:12).

But it was not only at the system level that concerns were raised. There was concern that the use of profiles would ultimately lead to the introduction of national testing.

The Australian experience was consistent with trends in school systems around the world where there has been a move towards the definition of learning outcomes and an emphasis on monitoring learning towards those outcomes. While techniques of monitoring are still based predominantly on the use of tests, increasingly teachers are being involved in the process of judging the achievement of students and reporting on progress, using direct observation as a major means of data collection. This will undoubtedly lead to a need for more defined, defensible measures of [learning] outcomes and to an increased interest in assessment methods. The sole use of standardised, norm-referenced assessment has become less acceptable as a source of data on teaching

effectiveness and on student attainment. Undoubtedly, the focus will shift more towards the skills and knowledge gained and on outcomes achieved. However, emphasis on assessment is misplaced unless an overview of outcomes-based education incorporates curriculum and instruction as well.

Consistent with this is the need to define clear statements of learning outcomes, and to do so in measurable terms. At a school or system level the outcome measures need to be macro in scope, for example, the numbers placed in employment, employer satisfaction with graduates, quality of course materials or parent or student satisfaction with their education. Specifically, detailed analyses of specific outcomes, as part of quality management methods and accountability to students, will increase. This will undoubtedly lead to a need for more defined, defensible measures of [learning] outcomes and to an increased interest in assessment methods.

Most states or territories have, by now, implemented some variation more or less derived from or related to the national profiles. Chart 1 below presents the extent to which the profiles have been taken up in Australian school systems.

Chart 1

System	Profiles Derivation
ACT	ACT Curriculum Frameworks (statements); Curriculum Profiles for Australian Schools
NSW	N/A
NT	Northern Territory Learning Area Curriculum Statements incorporating NT Profiles (statements and Profiles)
Qld	Student Performance Standards – implementation on hold (English and Mathematics profiles)
SA	Curriculum Profiles for Australian Schools
Tas	Curriculum Profiles for Australian Schools
Vic	Curriculum Standards Framework (Statements and Profiles) and CSF-II.
WA	Student Outcome Statements (Profiles)

In some cases (e.g. West Australia), the curriculum is defined by student outcome statements which in turn are defined as profiles. There appears to be an inconsistency of terminology and of documentation of the national curriculum. MCEETYA described the current state of implementation as a situation where almost all states and territories were using the statements and profiles, or local variations thereof, as a basis for their curriculum development. Western Australia, following a two-year trial of student outcome statements, was seeking ministerial endorsement to implement its version. Queensland, following a change in government, put on hold the implementation of Student Performance Standards pending a report of a review group. New South Wales, as a result of its 1995 review, was no longer adopting nationally developed statements and profiles for implementation through syllabuses. Syllabuses are instead being developed with outcomes related to NSW's five stages of compulsory schooling. However, the report of the review group recommended that syllabus support documents show how student work in stages could be linked to national profiles.

Griffin and Smith (1996a) reported that systems have modified the national statements and reduced the number of outcomes or levels involved. Western Australia has linked the learning areas and reduced the distinct identity of traditional discipline areas. The

Curriculum and Standards Framework-II is a further refinement of Victoria's response to the call for national statements and profiles. Common benefits were identified at a system level. For example, the national profiles and outcome statements gave everyone:

- a common language and framework for planning, assessing and reporting;
- fields of study defined;
- a national set of standards;
- a more transparent approach to assessment.

There were, however, concerns about:

- the complexity and volume of the documentation;
- a lack of consistency across and even within some statements and profiles;
- perceived differences in learning outcomes expected of students across learning areas;
- the fact that the lower level outcome statements were considered to be too complex.

All systems reported benefits in terms of:

- pursuing an outcomes based approach to education, and moving away from the inputs and process approach;
- stronger links between the teaching and learning, and assessment and reporting;
- explicit achievement statements about the students' learning for parents, teachers and students;
- the availability of data on the "health" of the system;
- guidance for future planning, programming, curriculum review;
- allocation of resources to match learning needs;
- enhancement of teacher professionalism through validation of judgement;
- expansion of the teachers' understanding of the field of learning and of expected student progress;
- collaboration between and within school communities in the fields of assessment and reporting;
- parity of esteem of curriculum areas.

Difficulties associated with the introduction of an outcomes based approach included:

- the workload generated for teachers in record keeping and assessment. Technology for managing this aspect of outcomes based education was considered essential, but this in turn became overbearing;
- costs to systems of professional development, materials development and time release;
- the workload for primary teachers facing change in all eight learning areas;
- suitability of the language of profiles for reporting to parents. A national project involving three states (Qld, SA and ACT) rewrote the statements in a language more friendly for teachers;
- Teachers' capability to make reliable assessment judgements without some form of moderation.

Most systems set the goal of phasing in outcomes or their equivalents over several years and all expected to have the curriculum in place by 2000. Most introduced assessment of English and Mathematics first. Some systems used the profiles for curriculum planning and programming.

The initial reactions to the use of profiles seemed to be ascribing too many difficulties or an extraordinary number of benefits to the implementation of profiles. Some of the potential benefits must have existed in the system before. Perhaps the implementation of profiles brings difficulties into sharp focus and poses possibilities as expected benefits.

The implementation of profiles required a shift to outcomes-based education, which, in turn, required a change in approach to and emphasis on teaching and assessment that schools were not able to accommodate fully. It requires a rethink of assessment strategies to recognise outcomes, and a change in teaching to emphasise students' performances. The first of three basic changes is the approach to assessment. This is associated with changes in teaching and curriculum development. The link between instruction, assessment and curriculum needs to become the basis for change.

The three components need to be integrated in a way that exemplifies outcomes-based teaching and assessment. This requires a shift from a transmission mode of instruction to a process and facilitation model that focuses on students' achievement of outcomes, rather than on relative performances and places, along with more emphasis on student responsibility for performance and learning, and less on the teacher as the main holder of expertise. In brief, nine explicit changes can be identified in the shift from transmission mode to outcomes-based training programs. The chart below illustrates the changes and links.

Chart 2

Shifting Teaching and Learning to Outcomes-Based Programs

Components and Change Criteria	Outcomes-Based Programs	Non Outcomes-Based Programs
1. Outcomes	Specific and measurable	Non specific, not necessarily measurable; typically global statements
2. Content	Outcome and student performance based	Subject matter based
3. Time	Student continues until the outcome is demonstrated	Fixed time units (semester, term) determine progress
4. Mode of Instruction	Teacher is a facilitator of learning using a variety of instructional techniques and groups	Emphasis on instructor as a transmitter of specialised information. Whole class is common
5. Focus of Instruction	What the learner needs based on outcome demonstration	What the instructor is able and likes to teach
6. Materials	Variety of text, media and real life materials based on various learning styles	Single source of materials (text or work books)
7. Feedback	Results are reported to the student immediately after performance in outcome-acquisition terms	Delayed feedback, usually in the global terms
8. Assessment	Criterion referenced assessments measure progress in terms of outcomes on learning continua.	Norm referenced assessments based on relative performance of others.
9. Reporting	Developmental and progress oriented	Grades, marks, scores and averages

In moving towards outcomes-based education, a fundamental change is required in the field of assessment. We need not only to change the nature of assessment, but also to change the language of assessment in order to bring attention to what is important. Norm referenced interpretations and reporting assessments in terms of grades, norms, means and so on are antithetical to outcomes-based teaching and learning. So too are other approaches that emphasise relative differences among students or distributions of scores, rather than the specific outcomes achieved by the students. Assessment needs to be developmental and emphasise the growth of competence towards defined outcomes.

The Language of Assessment

Assessment is an interpretive process because it involves a specific language which we use to construct an interpretation of our observations. We construct an interpretation of our students' learning based on the many inputs they provide for us. These inputs come in the form of pieces they write, their responses to texts, various assignments and projects they complete, contributions they make to discussions, observations of their behaviour in different settings, questions they ask in classrooms, interviews, performances or demonstrations involving application of specific knowledge, and specific tests of competencies. Observations are of the everyday behaviour and discussions exhibited in

classrooms. Teacher observation, judgement and interpretation must become an indispensable assessment process in an outcomes-based learning context.

When teachers assess students' work they make comments about what the student produces. The language used by the teacher indicates what is important and what is valued. More importantly, the language of classroom assessment becomes the language through which the students evaluate their own learning. If the language implies that there is one correct interpretation, response, behaviour, performance, or piece, the students may come to assess their own interpretations as either correct or incorrect and may not value diversity in the classroom. This is exacerbated at report time in schools. When teachers write reports, they have to find the words to describe a student's development in all its complexity. They are usually working within severe time, space, and format constraints. They have to take into account the diverse cultural backgrounds among the parents, students, and others that might read the report. Often, they are faced with reducing extensive and complex knowledge about the students' development to a single word or letter grade. This situation means that the greater the knowledge a teacher has of the student's learning, the more unreasonable is this task. Moreover, the communication produced is likely to be misinterpreted by the audience, as the report is reduced to symbols such as letter grades, percentage scores or short cliché comments out of context. Where formative application of assessment information is required, this is counterproductive. Reduction to symbolic reports without a commonly understood frame of reference leads to misinterpretation.

It is not just classroom assessment that is interpreted without the aid of an appropriate frame of reference. The community and the education system assess students, teachers and schools from the data they receive. Parents make sense of a test score or a report form or comment based on their own schooling history, beliefs and values. Assessment and reporting codes such as a test score or a letter grade, for example, may look "scientific" and "objective", but they too must be interpreted by the reader and this is always a "subjective" and value-laden process. The receiver's personal history and experience in assessment and schooling may be the frame of reference used to interpret the codes describing a student's performance in school.

It is not only the outcome or report of the assessment that is discussed. The process of assessment is also subject to discussion and the language of that discussion is also important. For example, teachers' professional judgment or direct classroom observations are often described as "informal" and "subjective" and contrasted with tests that are "formal" and "objective". Herste (1992) put forward the interesting point that a teacher's judgement is described as gut-feelings, or guesstimates. If the interpretation was derived from a view that held teachers' observations to be "direct documentation" and test results to be "indirect estimation", the interpretation of assessment data may be different.

Teachers develop an immense store of experience in the classroom and eventually automate their judgments. Their interpretation of what is seen in the classroom becomes intuitive and difficult to articulate. This difficulty has become a problem for the teaching profession, as outside observers interpret it as a weakness and tend to resort to more "objective" indirect and standardised means of making observations. What is needed is a series of direct observations and a standardised way of verifying these, then communicating them to a range of audiences.

A further difficulty is based in the terminology used for assessment. The terminologies change and different groups use them for different purposes or as situations change. For example, the term "norm-referenced" once meant that assessment data, and typically test data, were interpreted by comparison with other groups of students who were similar in some defined way to the students being assessed. The term is normally associated with standardised tests. A norm-referenced interpretation of a student's work might describe it "as good as that of 20 percent of the children that age in the country". Similarly, "criterion-referenced" assessment once meant simply that a performance was interpreted with respect to a particular level of performance—it met the criterion or it did not. In the earlier (and still common) meaning, most criteria for criterion-related tests were arrived at by finding out how a group of students performed on the test, then setting criteria in accord with what seemed a reasonable point of passing or failing.

The present understanding of criterion referencing is far removed from this practice. Norm referenced and criterion referenced interpretation of assessment information is better understood. For example, the discussion no longer refers to norm-referenced and criterion-referenced tests. It is more correct to discuss norm- or criterion-referenced interpretation of assessment information or of test scores. The emphasis is on the interpretation rather than on the instrument. This is an important distinction that has implications for assessments.

Teaching to tests or reports can also have both positive and negative implications for the classroom. Teachers required to improve test scores have been induced to "teach to the test". What is in fact happening is that the teacher is teaching to the score-based report that arises, because there is a feeling that the teacher is accountable for the score. If the teacher is accountable for outcomes achieved, the washback of the reporting process may be quite different. Now the pressure is not to improve a score, but to increase the likelihood of students demonstrating a larger range of definitive learning outcomes. Outcomes-based education encourages this form of washback from assessment and deliberately links the curriculum, assessment and teaching. The teaching and learning program can then become assessment and report driven.

In any teaching and learning program, a narrow focus on the type of information collected would be counterproductive. Teachers automatically have access to a whole range of information about students, and all of this is taken into account when making decisions about overall student progress and development. Customarily, a teacher has access to a range of information including data on attendance, scores on tests, etc., achievements, attitudes, motivation, learning strategies, personal matters, rate of progress and so on. However, systems also need information about student learning in order to make decisions about resources and policies and to monitor the health of the system itself.

There are differences between schools and regional areas and even between systems. Monitoring schools in terms of retention, transfer or progression from one level to another can be through indices or through the percentage scores on tests, but what is actually learned by these students is a much more important measure of the productivity of the school. Governments have increasingly requested an objective measure of this learning as a measure of outcomes. The more the schools achieve autonomy, the greater the pressure for accountability in terms of learning outcomes.

Developing Assessment in Schools

Assessment is being defined and redefined as new consumers of assessment information recognise its importance and become increasingly involved at levels of policy, development, practice and consumption. This is an era when assessment has broken through the boundaries of the social institutions formally charged with its application in education and psychology. It has been taken up by business and industry in ways that we had previously not regarded as possible. Until recently, when we spoke of assessment we referred to those practices that were concerned only with the measurement and quantification of learning in schools, colleges and universities. Educators and psychologists were the sole purveyors of techniques and instruments. They were the only ones recognised by society as able to interpret the meaning behind students' work in producing essays and completing tests, performances and other means used to identify how much an individual knew, could do, or needed to learn. The diversification of assessment practitioners and policy development means that many of the older understandings of the subject, so strongly founded in psychometrics, may be undergoing fundamental change.

I would like to draw out some of the ways that assessment has changed, and ways that it has stayed the same. We shall examine how its practices have spread from the classroom to the workshop, the corner shop, the office and the street; and how teachers and psychologists are now accompanied by specialist assessors, office workers, truck drivers, farmers and specialists from every area of recognised skill.

These new exponents bring with them new insights, approaches, beliefs and values associated with the assessment of learning. Emphasising the role of assessment in the workplace opens new opportunities for observation, interpretation, record keeping and decision making associated with it. More importantly for the educators and the traditional exponents of assessment, it brings a new and critical range of experts and practitioners who argue for direct and objective evidence of successful learning or development.

Outcomes-Based Education

In recent years, education has seen attempts worldwide to shift the emphasis from curriculum content to learning outcomes. Education and learning have changed their focus to what is learnt rather than what is taught. This is accompanied by the major problem that the nature of an educational outcome for school students has no generally agreed operational definition, and—perhaps as a consequence—the way in which many outcomes are defined makes them difficult to assess. In New Zealand it is even claimed that the outcomes are deliberately written so as to ensure that the interpretation is flexible enough to allow for individual and community differences. This is not surprising, as there are few operational definitions of outcomes, despite a literature devoted to the topic. They are nevertheless integral to assessment, and we will begin by establishing what we mean by an outcome. Jasa and Enger (1994) and Willis and Kissane (1997) synthesised the literature and listed several characteristics. They suggested that outcomes should be:

- descriptive of **long term** learning;
- **generalisable** to standards beyond the classroom;

- ❖ **developmentally** appropriate;
- ❖ observable and **assessable;**
- ❖ essentially demonstrative of **changes** in the learner;
- ❖ able to address **higher order** thinking;
- ❖ able to acknowledge different **learning styles;**
- ❖ **understandable** to stakeholders.

Yet this list still describes more of a philosophical stance on the concept, rather than what is expected of students. For assessment, we need to understand how to focus on evidence of change that indicates development—i.e., that is deeply situated and generalisable.

Perhaps we can borrow from the workplace example and extend the definition of outcomes to help with assessment. In the table below, five components of competence used in Australia can be seen to have parallels in outcome attainment. In demonstrating such attainment, a student is expected to change, link to other knowledge, demonstrate higher order knowledge, develop or build on previous skills and knowledge, and generalise or transfer the knowledge into new contexts.

Chart 3

Competencies and Outcomes: Components for Assessment

Workplace Competence	Educational Outcome
Perform individual tasks (task skills)	Speak and write, make and do things to show change as learners
Manage a number of different tasks within a job (task management skills)	Link to and incorporate into related skills and knowledge
Respond to irregularities and breakdowns in routine (contingency management)	Solve problems involving higher order thinking and deeper learning
Incorporate into the job/role, including working with others	Build on and incorporate into previous knowledge and skills base
Transfer to other environments and contexts	Transfer into new contexts

Each of these definitions has important implications for teaching, learning, assessment and curriculum development. They emphasise the importance of transfer and generalisation beyond the immediate context. Building assessment programs in and for schools therefore needs to be more than recall of factual knowledge, or the ability to write an essay, or select the correct alternative. Setting assessment tasks becomes more complex than preparing a simple item on a test.

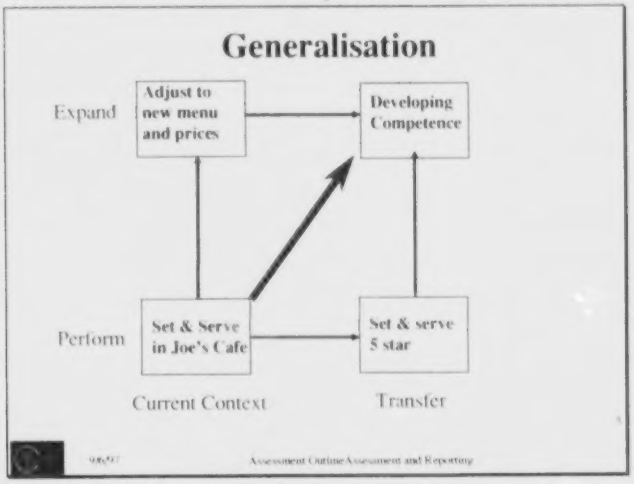
The two things that underlie these definitions, and are most important, are transfer and links to previous knowledge. Both indicate growth as a learner and give a direction to that growth. Without this underpinning emphasis driving assessment and learning in outcomes-based approaches to education, assessment becomes deficit-driven, norm-referenced and devoid of interpretation for individual student growth. Learning becomes dominated by what the assessment shows a learner *cannot* do. Implications of the outcomes-based education approach and generalisations of learning for assessment include the imperative that the demands on the learner must include contexts that he or she has not previously encountered. Moreover, the assessment tasks must fit into a

pattern of development and growth that give an indication of past and possible future learning relevant to the set of outcomes and their interpretive framework. Assessment takes on the role of identifying thresholds of performance and knowledge in the individual's development. It does not provide an exhaustive checklist of isolated achievements and deficits: the former is the equivalent of driving by looking through the windscreen, and the latter of driving by looking in the rear-view mirror. One of my assessment colleagues in the hospitality industry showed me this with an example of a waiter setting tables in a café. If only a list of things completed is used, the transfer and development aspects of the competency are ignored.

This is shown in Figure 1. It illustrates the concept that although the competencies of table setting apply across two domains, the two contexts add new conditions to the assessment. Thus narrow lists of achievements would not enable any generalisation to be made. Neither would this assessment on its own. The idea of assessment for prognosis rather than diagnosis emerges from this approach. The potential for further learning, rather than a need for catch-up training, becomes the focus of the assessment.

Figure 1

Assessment, Development and Context

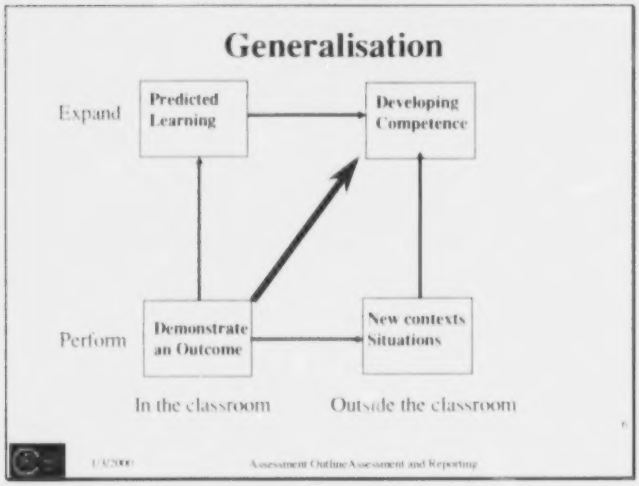


These same principles apply to the classroom. Figure 2 illustrates how these ideas combine in a criterion-referenced model.

BEST COPY AVAILABLE

Figure 2

Outcomes Assessment as Generalisation



Developmental assessment depends on establishing progressions of learning outcomes as a basis for interpreting assessment information. Where curriculum frameworks have been developed, and profiles of learning exist within those frameworks, the hard part is done. The developmental progression has been largely identified.

Developmental Assessment

Developmental assessment does not underpin anything dramatically new. Bloom (1956) proposed developmental learning and assessment. Gagné (1965) modelled hierarchical progressions of learning. Rasch (1980) modelled it with mathematical precision. Glaser (1981) gave it a name and Vygotsky (1962) showed how progress and development can be understood and encouraged. Biggs and Collis (1989) developed a means of understanding the outcomes placed along the progression, and these influenced the development of subject profiles in the late 1980s (Griffin & Jones 1987; Griffin 1990). These are important underpinnings of assessment, because we are always examining learning, and to separate our understanding of what is learnt from how it is learnt is fraught with dangers of trivialising the assessment and encouraging surface and shallow learning and performance. If, in assessment, we ignore how learning occurs, then the tasks used are likely to become irrelevant to the competence or to the learning outcome. This has become a common criticism of tests as a means of assessing learning outcomes.

Vygotsky's theories have been applied in adult learning, and Bloom's work was used in vocational assessment with the classification of knowledge, understanding and application. Few extended the taxonomy beyond that, and Gagné's work has been used in the definition of training and development of learning materials for industry sectors. These theories have had profound influences on our understanding of learning and assessment for more than thirty years, but they may yet be jettisoned in large-scale assessment programs. Unless assessment is linked to learning, and its design is underpinned by an understanding of how developmental learning occurs and is assessed,

there is a distinct danger that assessment of outcomes will regress to the models of the 1970s and will suffer the same fate. I will take up these points later in the discussion.

Assessment Defined

Assessment is an emotional subject. Threats and fears abound and often obscure the logic and well established theoretical models of learning and measurement. I should describe what I mean by the term assessment and perhaps encourage a "back to basics" movement in assessment rather than learning. About fifteen years ago, my colleagues and I were working with the Victorian Department of Education; we enjoyed positions as researchers in a Centre for Research in Evaluation, Assessment and Measurement. During the course of a curriculum-renewal project for the education system, we became aware that assessment never occurred without someone being informed of the result. Reporting was inseparable from assessment. While this seems obvious now, and the two words seem to be inseparable when referring to school-based assessment, it was not so clear at the time. We searched the literature for references to assessment and reporting. There were many linking measurement and evaluation, assessment and measurement, and assessment and evaluation; there were some on reporting, but none on assessment and reporting. This fascinated us. As we observed the practices in classrooms and in schools, a view emerged that has to this time remained unchanged. We observed teachers monitoring learning as they taught and reasoned that it was a continuous, intuitive process that never stopped during the course of teaching and learning. In fact, assessment and reporting was identical to teaching and learning. It stimulated us to think about why an assessment was being conducted, and the implications of those reasons.

It was clear that assessment was not always a systematic process. In many instances it was ad hoc, informal, intuitive and unstructured; in others, it was a highly structured ritual, with definite roles and relationships among the players. At the heart of our assessments we collected information, interpreted it and made decisions and inferences about the capabilities of the students. The more systematic forms of assessment involved more formal records and communication between stakeholders—even a teacher's observation of a student involved all these components. When it was all put together, assessment and reporting became for us *a purposeful process of gathering, interpreting, recording and communicating inferences about learning to stakeholders* (adapted from Griffin and Nix, 1991). Each of these components can be a field of study in its own right. Studies of purposes, procedures, methods, interpretation, record-keeping systems, communication modes, inferential patterns and procedures and stakeholder roles, rights and responsibilities have emerged. Not all of these are technical, nor do they require the sophisticated psychometric and statistical skills that once separated (and perhaps alienated) the measurement and testing specialist from the practitioner. The components of assessment are being explored and developed and influenced by curriculum, policy, industry, entrepreneur and training specialists as well as by statisticians and psychometricians; the challenge is to draw from all their work and to improve the overall effectiveness of assessment. Some definitions have distinguished between evaluation and assessment by focusing on the level of data: assessment was regarded as being about the student and evaluation about the program. While this, in general, may be true, there are

many instances where the distinction breaks down and its usefulness is doubtful. Measurement has been defined as the assignment of numbers; evaluation has many definitions, but among the most common is the notion of judgement of worth. Assessment is separated from evaluation by the distinction between judgement decisions and inferences.

So much has been written about assessment in schools, in particular, that the subject matter can hardly be dealt with in a single presentation. The literature is distinguished by its history, its applications of mathematical modelling, its empirical base and its links to psychological theories of learning. Recent emergence of fuzzy logic and chaos theory in assessment may lead to new levels of understanding, given that decisions are made using unclear criteria. The history of assessment research is dominated by analyses and techniques of enhancing our understanding of tests, and in particular of multiple-choice tests. Possibly no other endeavour in education is so well understood by researchers and so thoroughly mistrusted by teachers. The irony is that so little is understood about teachers' inferences and judgements and so much is mistrusted by decision makers. Trust is not a characteristic of assessment programs. Research into the mathematical modelling of test-takers' behaviour has not resolved the lingering mistrust of the test and its uses. Studies of judgements have not resolved the mistrust of teachers' judgements. While it is possible to account for the minutest difference in amounts measured, we have not solved the huge gap in teachers' trust—nor have we resolved the gap in the administrators' trust of teachers' judgement. There is an urgent need to develop procedures that allow teachers to exercise professional judgement of learning in a way that brings an air of confidence to the outcome, or inference, of the assessment. The two components, judgement and trust, are inextricably tied together in assessment. They have traditional counterparts in the notions of reliability and validity. We will return to this later. The problem is unresolved in delivering a reliable assessment and an inferential process that can be trusted.

Understanding the process of assessment enables us to apply it to a number of contexts. The teacher asking questions in class is purposefully gathering information, interpreting it, temporarily storing it and responding to the students. A government setting tests for student cohorts gathers information, interprets it, records the data and communicates them to a range of stakeholders. I recently attended the school concert at my grandson's school. Here, in the form of a concert, was a school making a statement to the community about the music program, and about the children's capacity to perform tasks and entertain, work in groups and represent their school. It was a powerful way of reporting these attributes at a school level, of showcasing some special talents and of reporting to parents and the community. In the late 1970s my colleague, John Izard from the ACER, and I were studying apprenticeship training; we observed workplace supervisors checking on their apprentices. They judged when to intervene and when to leave alone—when the apprentice could be relied upon to perform the task or a set of similar tasks without supervision. When the apprentice could deal with problems, manage the tasks required and see how they fitted into the overall pattern of jobs to be done, the supervisor generally felt that he or she could be left alone and trusted to do the job. Such a scenario is repeated in every sphere of life; one group of people observes another and makes judgements and inferences about the possible independence of performance. The observer, the observed and the process of observation vary from

context to context, as does the need for records, judgements and skills in interpreting the procedure; but whatever the context, observations are made, interpreted and recorded, and a decision communicated from one person to another.

The language of assessment also is being redefined. We are becoming accustomed to the rhetoric of learning outcomes, competencies, learners, teachers, trainers (or providers) and, recently, "assessors" and "assessees". As we already have seen, there is an interaction between the one who does the judging, the one who is judged and the audience for the report. The audience must trust the judge to gather the right information, interpret it appropriately and then translate it into a communicable inferential message. The judge can be as remote as a government department or as close as the teacher (or even the students themselves). The judged can be a large cohort or an individual; the audience can be as diverse as the person being judged—a class group, the provider (or school), a profession, parents or employers. What must be common to them all is faith in the judgement.

I am extending this discussion to emphasise a point. Each of these examples involves a purpose, a process of data gathering, interpretation and communication. Assessment does not need an emphasis on the use of a test, on statistics, on measurement or evaluation; the utility of a generic definition has recently become more apparent as the emphasis on assessment in the workplace has increased. While emphasising assessment and reporting in schools and in industry, there is no implication that other contexts are unimportant. The developments in schools and industry, however, offer such exciting new approaches, roles and definitions that they will more than fill the allotted time to be explored today.

Purpose

Nothing is done without a reason. Nominating a purpose as a fundamental component of assessment does not imply that many assessments are capricious. Assessment purposes have been listed in numerous textbooks and papers as including almost all aspects of education and training. The Green Paper clearly sets out the proposed purposes of the national testing program.

In schools, the system and the teachers conduct assessments. Willis and Kissane (1997) describe potential conflict, in that teachers assess the *enacted* curriculum while systems assess the *intended* curriculum, but possibly neither assesses the *achieved* curriculum; an outcomes-based curriculum should, in theory, close this gap. If both the system and the teachers are attempting to help students achieve the same set of outcomes, the assessments should be outcomes-based.

To list every purpose that assessment has been given would be impossible. However, most purposes ascribed to assessment (not to measurement or evaluation) may have one common denominator. Consider this list:

- provide information to parents and show relative national performance
- identify strengths and weaknesses
- show individual learning needs
- moderate teacher professional judgement
- evaluate teaching and learning programs

- ❖ report on school performance relative to nation and similar schools to evaluate school programs
- ❖ provide early intervention for children needing it
- ❖ improve data analysis and strategic decision making
- ❖ develop policies and monitor their effectiveness.

The list has a common underlying purpose. Whatever the stated purpose, the exercise is expected to be predictive. This is not surprising, given the inferential and generalisable nature of assessment programs. Underpinning almost every assessment, regardless of whether it is for improvement (formative) or credentialling (summative), there is a need to report to an audience. For the audience, there is an expectation that the assessment will be predictive of future performance and go beyond the immediate content and tasks used. In reporting, we need to be able to describe what can be done in a range of contexts, to help predict what a student is likely to be able to do—or learn to do—in the future. Context effects are now considered not as "psychometric noise", but as part of the evidence used to generalise and predict. On this basis we make decisions about teaching, learning, training and all the associated resource implications; hence, predictive validity becomes the *sine qua non* of assessment. As Messick (1995) argues, validity is about inference and consequence. Without it, assessment is useless. Now, given the prognosis of further learning and future performances, the need to understand the developmental nature of learning becomes increasingly important.

The list of purposes for the New Zealand testing program is impressive in its scope, ambitious in its breadth and perhaps overstated. To demand so much from a single data collection exercise is optimistic. Improving policies, for instance, is surely dependent on having policy questions in the first place. Only policy aims occur in any of the documentation I have read. Policy questions, such as the effect of home background, school characteristics, teacher characteristics, teaching conditions and practices, or student motivation, are all relevant issues to address in such a study. Certainly there is no mention of how these questions will be addressed. Nor is there anything in the documentation about how diagnostics will be reported, or how school and classroom programs will be evaluated, apart from through comparisons to other programs. At risk of generating further controversy, the program would well do with additional data collection that enables the achievement levels to be linked to and explained in terms of questions such as:

- ➢ How is variation in achievement among regions, groups, and school types related to differences in curriculum?
- ➢ How can variation in achievement be related to variation in assessment practices?
- ➢ What school characteristics are associated with differences in school level achievement?
- ➢ How can schools improve their student skill level over time?
- ➢ How are the differences in achievement across grade levels related to practices, resources and school policies?
- ➢ How can these studies be linked to international studies of achievement and comparisons made to achievement levels at those times?

What would we make of the differences?

- Which subgroups are to be compared and why?
- How can malleable variables be used to identify intervention strategies for improving lower skill levels for some groups?

Ironically these need more rather than less data collection; but in doing so, they signal to teachers that the system is serious about identifying those characteristics that can be changed for purposes of intervention. Simple relative position of test scores engenders a guesswork basis for intervention and policy development.

Observing

So many methods of observing student learning and skill development exist. Essentially there are only four things that we can ask the students to do. These are to write, make, do or say something. Their actions must be observable and monitored in such a way that we infer covert development in such areas as cognitive and learning developments. The inference is the single most important aspect of assessment. We often forget this in the emotional and rhetorical arguments about how best to gather the evidence of learning. The argument focuses so firmly on how to observe (or not to observe) that any overall purpose is lost.

Tests are obviously the most used method and, of these, multiple choice tests are by far the most common. They need not be. The Maryland State Department in the USA involves teachers in clusters to design teaching and learning programs, and assessment is entirely performance based. The OECD PISA study is also developing new styles of tests that allow for much more flexible approaches to item development. In Tasmania we have (Callingham and Griffin 1999) shown how complex, problem-solving tasks can be used to identify broad thinking among students and that this can be monitored over time, even allowing teacher intervention by teaching. This was a feature of the Australian National Literacy Study, which also managed to incorporate student portfolios. The thing that is common to recent test development is the policy that the assessment procedures should enable students to take the opportunity to demonstrate abilities. This should be an integral aspect of the teaching and learning program. Treffers' (1987) distinction between construction by students and production against specific prompts is becoming a reality. There is so much that is possible. What determines the choice of approach is largely the expertise of the developers and the budget available for recording and analysis. A long term, take-home example of an investigative task was also used in Tasmania (Callingham and Griffin 1997).

The difficult part is the definition of methods of recording student performances, and designing these so that they can be interpreted in terms of knowledge and skills demonstrated but, more than this, so that they fit within an overall framework of development for the students. The tendency for many systems is to use multiple choice items. Criticisms of these are well known, but they do have strengths and can be used to identify skill development within and between students over time. The criticisms of these items also apply to other test item types. What is important is the information they provide and the capacity to generalise beyond the set of tasks obtained. People familiar

with the interpretation of factor analytic structures should understand this process. What all these examples illustrate is that the observation methods are varied and that the choice of approach depends mainly on convenience, expertise, cost, and practicability. All of the methods are capable of providing information that can be used to generalise and infer learning development. The emotive differences expressed regarding methods are based more in issues of face validity. The inferential validity of large scale testing programs, such as in TIMSS and other IEA studies, is well established and used by many to justify arguments for policy changes, resource allocation and changes in curriculum.

Recording

Recording observations is important, but it is usually neglected, and often misinterpreted, or taken for granted. Students responding to multiple choice items are recording their own performance. A teacher scoring a performance task or an essay is developing the record of the performance. We can maintain types of records such as samples of work (portfolios), and, more recently, complete large numbers of paper and electronic checklists. The latter have become the most prevalent form of recording. Unfortunately, the checklist is also seen as an assessment strategy (Docking 1996)! Victorian schools are now recording every "outcome" statement in the Curriculum and Standards Framework (Board of Studies 1995) in electronic checklists; this is unfortunate, because checklists tend to become the exhaustive, compulsory and exclusive list of what is important, and thus limit the opportunities for holistic, generalisable assessments. To over-emphasise them leads to a deficit model of learning, as teachers and trainers focus on what is lacking, or on gaps in the lists. Relationships between items on a checklist tend to be ignored and the notion of developmental learning is lost. On the other hand, checklists are important when we are observing procedures or performance demonstrations in which specific steps must be taken. For those purposes it is hard to improve on checklists as a recording device; they are, however, simple to use (simplistic) and overdone. They can be valuable, as we will see later, but their value depends on how the records are interpreted and on the inferences that are made.

Interpretation

To make inferences, we also need evidence and ways of interpreting it. Assessments have often been described as norm-referenced, or criterion-referenced. This is basically incorrect. There are, actually, no such things as purely criterion-referenced tests or norm-referenced tests. Referencing is associated with the way we interpret—or make sense of—the records of performance, not with the nature of the task performed, nor indeed of the performance itself. The referencing process is a step towards providing an inference. When we use tests we are usually trying to make sense of a number or score. Reporting just the score or a range of scores is not interpreting the test performance. In performance assessment, we have to interpret characteristics of the performance and rely on records of observations, sometimes in the form of checklists and at other times in the form of written statements. With portfolios and other forms of work-sample assessment we have to interpret evidence in the form of a collection of artefacts. In order to make sense of observations and use them for prediction, we use frames of reference based on three

things: the *group* to which the person assessed belongs, the *tasks* that are being performed, and the *individual performing or observing* the task. These are the bases of norm-referenced, criterion-referenced and ipsative- or self-referenced interpretation.

Norm-referenced Interpretation

Norm-referenced assessments are *based on expectations of the group* to which the individual belongs. Performances are regularly described and qualified in terms of what might be expected given the racial, educational or home background, physical or intellectual ability, command of English, exposure to a particular context or program, or amount of time spent on task. When the justification for a decision of learning is based on the nature of a group of people, a norm-referenced interpretation is used. In the quest to avoid norm-referenced comparisons, we excuse or praise a performance by describing how it accords with expectations, given the person's membership of a specific group or population—we hear lower levels of achievement being accepted because of location, language background or other group characteristics. We even rephrase our expectations to meet the different needs of individuals or groups. This is norm referencing.

Prolonged experiences condition our expectations of what is "normal" or average. The expectations soon become the standards; it is a common practice that is easily recognised. The irony is that local (or special group) norms are used to argue against standardised norm-referenced assessments. This is the case in the Green Paper, where exceptions are made for special groups.

Criterion-referenced Interpretation

Criterion-referenced interpretations focus on performance, not on group membership. The term is becoming more popular, but it is used to describe many things. In reality the use of criterion-referenced interpretation is quite rare.

The idea is not new. It has been around since the early 1960s and has been the guiding principle for one major era of education—that of competency learning in the 1970s. Unfortunately, during that period criterion referencing lost its way. We focused on behavioural objectives, and cut-off scores for mastery tests. Criterion referencing became interpreted as a means of determining whether someone had mastered specific objectives.

Glaser first defined criterion-referenced performance as assessment interpreted through the tasks performed. He clarified it later as "the development of procedures whereby assessments of proficiency could be referred to stages along progressions of increasing competence" (Glaser 1963; Glaser 1981). A large amount of developmental work is still needed to identify suitable progressions of increasing competence, although we as a nation have started to move towards such progressions via such frameworks as the Australian National Profiles.

Criterion referencing does, however, have its weaknesses. Progressions of increasing competencies can only be used effectively when both competencies and progressions are stable. This is not the case in many areas of knowledge and competence. In computing and information technology, for example, the continuum is rapidly changing, and it is likely that as soon as we could define the continuum and develop a test, we would need to revise it. This is true at the developmental edge of knowledge, but hopefully not at the

beginning or underpinning stages. In areas such as literacy and numeracy, it is arguable that the knowledge base is sufficiently stable for criterion referencing to be successful.

Ipsative-referenced Interpretation

Ipsative referencing requires observation of an internal set of values, preferences and attitudes. In assessment, this has recently become important again with the shift towards observational procedures, judgement and rating-scale forms of assessment. But it is the assessor, not the assessee, who makes the judgement, and he or she is expected to make it independent of the presence of other information, personal preferences, values and perhaps even prejudices. We know, however, that these can and will confound the judgement. The influence is more pronounced when the criteria are not explicit (Messick, 1994); but even when explicit criteria are provided, other pressures can influence the judgement considerably. When the stakes are high, judgements are particularly vulnerable to personal relationships and an awareness of consequences for both the assessor and the assessee (Messick, 1994). My colleague here at the University of Melbourne has recently demonstrated this in her investigation of workplace assessors (Gillis et al., 1997). Comments on the Green Paper have lamented the absence of monitoring aspects of values and attitudes, but give little advice regarding the methods of collecting and have studiously avoided articulating any possible inferential basis of results that might be obtained. Governments and industries shy away from these measures as they are so hard to enforce. In California, they have led to the demise of entire curriculum programs in the state's school system. There is no doubt that monitoring these things is going politically where angels fear to tread.

The influence of personal factors on judgements is technically called bias if the interpretation is not directly related to the criterion of interest. Gillis et al. (1997) showed that in workplace assessments, adjustments were made for language, appearance, age, gender and perceived ability, and that these adjustments differed for on-job and off-job assessments. Being flexible was interpreted as making adjustments to standards, rather than to the process. One hopes that this is not the case in the Green Paper discussion of flexibility. The use of experts in a particular curriculum area indirectly endorses connoisseurship in assessment, in which an expert's intuitive knowledge is used to make fine distinctions between performances (Eisner, 1993). This is relevant when the interpretative frame of reference is directly related to the assessor's own personal experience and expertise, in terms of quality, experience with the task, criteria, and understanding of the expertise required and the connoisseurship involved. It is regularly used in the assessment of written language, and in second language acquisition, in the assessment also of spoken language.

What are Criteria and Criterion-referenced Assessment?

In criterion referencing, lists of outcomes are necessary, of course, to communicate to stakeholders. However, the generalised use of checklists of achievement and performance is more in line with the corruption of criterion-referenced approaches of the 1970s and 1980s (Popham, 1978 and Glaser, 1963; 1981). In Victoria there is appalling, mandatory use of a computer software package called Kidmaps, which requires teachers to rate every student on every outcome statement. The emphasis on checklists, and atomistic

approaches to every outcome, indicates a confusion of mastery learning with outcomes assessment. The emphasis on mastery learning is exacerbated by compulsory use of checklists. This can lead to a breakdown of an outcomes-based approach to education (Slavin, 1994). Perhaps both pre-service and in-service teacher education may need to shift to emphasise outcomes-based education, the role of assessment and criterion-referenced interpretation in particular.

To fully implement Glaser's (1981) notion of criterion-referenced interpretation, a number of thresholds are needed, each representing the division between increasing levels of competence. Checking where an assessee is in relation to a series of thresholds does several things. First, it illustrates what has been achieved so far; second, it illustrates what the assessee can do now; third, it forecasts where the progress in learning is likely to lead. Using competency assessment to link only training for non-competence (or "cannot do now") is a deficit model, which, while important, is of limited use. The "can do now" information should enable us also to forecast laterally into new contexts and different learning outcomes, and give breadth to the interpretation.

A criterion describes the point at which we decide something changes from one state to another or moves from one level to another. Criterion referencing is the process of referring our evidence to this threshold in order to decide which side of it the observation best matches. Thresholds, however, are notoriously difficult to define or describe.

Some years ago, Rowley (1982a, 1982b) illustrated how difficult defining the thresholds can be when he wrote a parody assessment, focusing on the question of beardedness. When do we decide that a man has a beard? How long do his whiskers have to be before we stop describing him as unshaven and start describing him as bearded? Here are several examples of these transition points or thresholds:

- Not (yet) competent | Competent | Highly competent
- Honors | Credit | Pass
- Level 1| Level 2| Level 3| Level 4| etc. — (Australian Qualifications Framework) (National Profiles)
- Clean shaven| unshaven | stubble | designer beard | bearded | full beard

According to Eisner (1993), the use of criteria involves an exercise of judgement, the provision of reasons for the judgement, and an understanding of the relevance of criteria to the genre assessed. The descriptions that national profiles provided fulfill all these requirements. While this may seem difficult to achieve, benchmarks and competency standards provide excellent vehicles for developing appropriate National Profile forms of criterion-referenced interpretation.

Standards and benchmarks

No discussion of criterion referencing is complete without a mention of standards. Standards for today won't be acceptable in ten years' time. The threshold we use to make a distinction between "adequate" and "unacceptable" is established using the experience of a group and an understanding of what can reasonably be expected. It is also a norming process, based on an understanding of the task itself—and it can change, as can the resources that may be used.

Standards are not fixed. They are commonly adjusted as we seek better or higher quality outcomes. The very fact that we acknowledge improvement of standards means that we implicitly acknowledge the existence of the underlying continuum that allows a standard to move; standards, therefore, are defined as one threshold or a continuum. In a competitive environment, we monitor each other's performances and identify examples where the standards are higher than those we set for ourselves; this is a process of benchmarking, where we observe others' standards in order to adjust our own. A benchmark, then, is a standard or threshold to which we aspire. This is sometimes called a "stretch benchmark".

When we understand how and why standards change, we will be in a better position to pre-empt these changes in developing better teaching and learning. Willis and Kissane (1997) (Sadler 1987), in arguing for standards-referenced assessment, lost sight of the relationship between criteria and standards. Standards referencing is simply a subset of criterion referencing with one threshold on a continuum with two levels.

Inference

Inference is the heart of assessment. At this stage the interpretation of all the evidence is synthesised and a single decision is made. Validity is now described as referring to the *use* and *interpretation* of the *evidence* collected, as opposed to the assessment method or task (Bennett 1993, Linn 1994, Messick 1992, Wilson, Scherbarth, Brickell, Mayo and Paul 1988). An assessment task that has been developed and endorsed for one specific purpose and target group may not necessarily be appropriate for another purpose or target group. Cronbach (1971) argued many years ago that assessors do not validate a task/test, but instead they validate the interpretation of the assessment evidence and the inference made from it. *Validity is a property of the inference*—not of the method of gathering data, or of interpreting it (Messick 1989). Consequently there is a need to define the construct being measured and be clear how the evidence relates to that construct. Moreover the audience needs to be able to see how the evidence provides the kind of generalisation and inference communicated to them.

Validity of assessment is inextricably linked to the definition of standards and criteria as well as the use and consequences of assessment decisions (Messick 1994; Gillis et al. 1997). There are many considerations for those responsible for developing assessments, especially when the consequences for individuals, for enterprises and for the national goals of education and training are considered. Assessment training for teachers is often token, and sometimes limited to a few lectures in undergraduate courses. Under these circumstances it is not surprising that a lack of trust persists with teacher assessments. Validity of a teacher's judgement cannot be assumed. To be valid, assessments need much more than just the simple, direct unstructured observation of performance. The need for expert assessment skills in schools must be taken seriously. Untrained people cannot expect to be specialists in assessment, and therefore cannot expect to bridge the gap in trust of teacher judgement that causes so much anguish.

Like the outcomes teachers are expected to assess, the school-focussed assessment standards are mostly couched more in terms of principles and philosophy than of

performance standards or competencies. They are easy to agree with, but difficult to translate into assessment practice.

This is interesting, given that the introduction of subject profiles in Australia and of developmental assessment world-wide has led to new forms of assessment and reporting, and of teaching styles and curriculum development. Classroom practices have shifted from a testing and psychometric base to a holistic judgement and descriptive base. In the light of these shifts, Harste (1992) proposed the following standards. They read well, but may be difficult to achieve without a deliberate effort to train teachers in their meaning and application. Intuition and good will is not enough, despite its copious supply in the classroom.

1. Educational assessments must specify what students should know and be able to do. The standards should be clearly defined before assessment procedures and exercises are developed.
2. The primary purpose of assessment systems should be to assist both educators and policymakers to improve instruction and advance student learning.
3. Assessment standards, tasks, procedures and uses should be fair to all students.
4. The assessment exercises or tasks should be valid and appropriate representations of the standards and outcomes students are expected to achieve.
5. Assessment results should be reported in the context of other relevant information.
6. Teachers should be involved in designing and using the assessment system.
7. Assessment procedures and results should be understandable.
8. The assessment policy and practices should be subject to continuous review and improvement.

There is, in addition, an international set of standards for measurement and assessment that address codes of ethics for examination and monitoring authorities, but not for teachers. It seems improbable that despite the development of competencies for teachers, only limited references are made to assessment and reporting. Even more surprising is the fact that assessment components in teacher education are limited to a few token lectures. I am currently in my fourth teacher education institution and over the past 25 years—my time in teacher education—assessment preparation has diminished from a pronounced emphasis on psychometrics and multiple choice test development to the present curriculum emphasis. These shifts have coincided with a strong rejection of testing programs and a move to more qualitative and post-modern paradigms.

Recent returns to testing and outcomes-based education have caught the teaching professional unprepared, untrained and vulnerable to assessment programs. No doubt the link between curriculum, instruction and assessment is real and an understanding of the specialism of the curriculum content is essential. However, it is clear now that skills in assessment design, especially those associated with developmental learning, have all but disappeared from teacher education and from the profession. When formalised assessment projects are undertaken by systems, for schools and for governments, the dearth of design skills retards real development. The situation is so serious that one Australian state system recently was moved to purchase test items from a United States testing agency to fill a void and use those items to market its online assessment to other states and countries. The job of scouring these to develop an item bank for an Australian system

and to ensure that they match the local curriculum is a task that needs to be done. In addition, it will need to be *seen* to be done by teachers, to make sure that that trust (validity) is not undermined. If it is not done, of course, we are likely to be looking for improvements against a foreign curriculum in Australian schools. The damage this could do to trust may be irreparable.

It can be argued that the demise of the multiple choice test question in classroom use has made the skills of item design redundant. This is not so. Good design is based on principles of assessment, evidence collection and interpretation. There is an unfortunate tendency for designers to use constructed response items, portfolios or direct observation and to replace tests with checklists. Interestingly, the same criticisms (atomistic, reductionist, shallow) that were used to label test designers in the past now apply to the judgement-based assessors relying wholly on checklists, now mandatory in Victorian schools. Supply-type items are easy to set, but the recording and scoring rubrics are much more difficult. Often this side of recording assessment is ignored and poorly designed supply or construction tasks are preferred because they look like classroom exercises. Difficulties of reliability (and hence validity) are often glossed over.

Reporting

Consideration of the way information is used is a central issue in revised assessment practices. It is necessary to consider how to make the information available to parents, students and teachers as well as to systems and administrators. All these stakeholders need information so that they can effectively plan instructional programs and improve learning. This is one area of assessment that offers a great deal of room for improvement.

Browne (1997) put this much more forcefully. She argued that the process should involve the following steps:

- State the learning outcomes, their sequence and demands
- Design the assessment and define the ideal response;
- Develop the teaching and learning procedures to achieve those skills outcomes
- Determine which outcomes are useful to various user groups
- Report on the outcomes achieved to each of the user groups.

Of course this is controversial, mechanical, robs schools of professional freedom and steers the teacher on a path towards changing practices, but it provides the information for which teachers are very much accountable to students, to parents and to their school. Accountability of a teacher to a system is a non-intervention process. Alternatively the accountability of the system to the teachers is a serious issue when wholesale evidence arises that skills are not developing and intervention becomes a political imperative. This appears to be the case now in New Zealand. The response is to change the assessment.

The Green Paper outlines an intention to include teachers in the assessment models, but there is no mention of the ways in which they will be trained. Will it be by osmosis in the development of assessments? Their assessments will be moderated against the national testing programs; but these programs are to produce norm-referenced distributions. This will induce teachers to worry about differences in distributions without any idea of what to do about it. This is teasing the education system, not testing

it. It is counting, not accounting; a stick without carrots, leading from behind in the hope that the teachers will find a way. The assessment programs may make many changes at a system level, but their chance of changing programs at the classroom level is minimal. Teachers will be provided with formats of assessment, and some new but familiar testing procedures, but this reduces chances to alter the communication between teachers and students and between teachers and parents. Nowhere in the Green Paper is there any reference to training teachers to communicate with parents through a parent-teacher interview, or to communicate student developmental progress in the existing school report format and the "book of school report phrases for all occasions".

It would be cheaper to conduct teacher communication-training programs, emphasising how reports can be based on records of direct evidence of learning (such as portfolios) and anecdotal evidence. The skill of linking these into an overall judgement of performance and an inference of progress in a learning construct is not a trivial one. Emphasising such a skill for teachers, and holding them accountable for the accuracy of these reports, would perhaps pay great dividends in the classroom and maximise the chances of improving student learning. We would still need the skills of the psychometrician to provide the interpretative frames of reference.

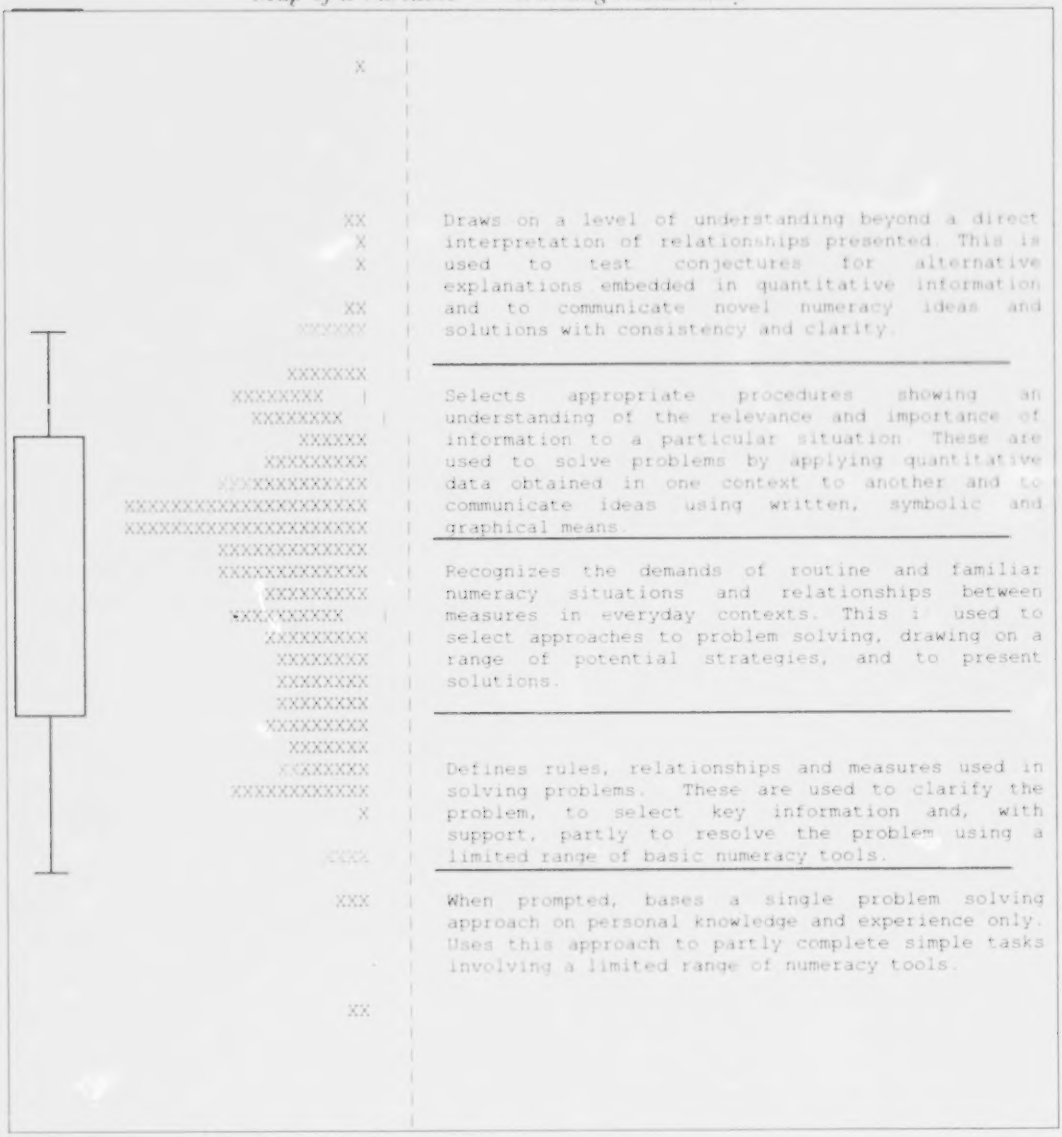
For example, in a recent assessment program in Tasmania, a performance task was delivered to schools via the internet. Students responded to the task of analysing a newspaper using a series of computer spreadsheets; they wrote their explanations and attached charts and other files to complete reports that were emailed to be centrally marked (Office for Educational Review 1997). A team of raters was used to score the student work, using a series of descriptive five-point scales applied to eight components of the task.

The eight components were combined with items from paper and pencil tests and scaled using a Rasch partial credit approach (Adams and Khoo 1995). Figure 3 shows the distribution of students on the left (x's represent students) and on the right, the description of the student performance and the score assigned. The central axis simultaneously presents both student ability and task difficulty, and is analogous to the continuum of increasing ability. Our job is to refer the student work to levels on this continuum, but first we need to define the levels of increasing competence.

Decisions about where to place the criteria or thresholds in the scale are yet to be technically resolved. We established a series of ad hoc rules for identifying the breaks (criteria) between levels, but these rules remain the focus of continuing study.

Figure 3

Map of a Variable — "Working Numerately"


Draws on a level of understanding beyond a direct interpretation of relationships presented. This is used to test conjectures for alternative explanations embedded in quantitative information and to communicate novel numeracy ideas and solutions with consistency and clarity.
Selects appropriate procedures showing an understanding of the relevance and importance of information to a particular situation. These are used to solve problems by applying quantitative data obtained in one context to another and to communicate ideas using written, symbolic and graphical means.
Recognizes the demands of routine and familiar numeracy situations and relationships between measures in everyday contexts. This is used to select approaches to problem solving, drawing on a range of potential strategies, and to present solutions.
Defines rules, relationships and measures used in solving problems. These are used to clarify the problem, to select key information and, with support, partly to resolve the problem using a limited range of basic numeracy tools.
When prompted, bases a single problem solving approach on personal knowledge and experience only. Uses this approach to partly complete simple tasks involving a limited range of numeracy tools.

There are several steps in the interpretation process. The box-and-whisker plot on the left establishes norms for the group, but fails to communicate any substantial information. The box represents the middle 50 percent, and the whisker lines below and above extend to the 10th and 90th percentiles of the distribution of students.

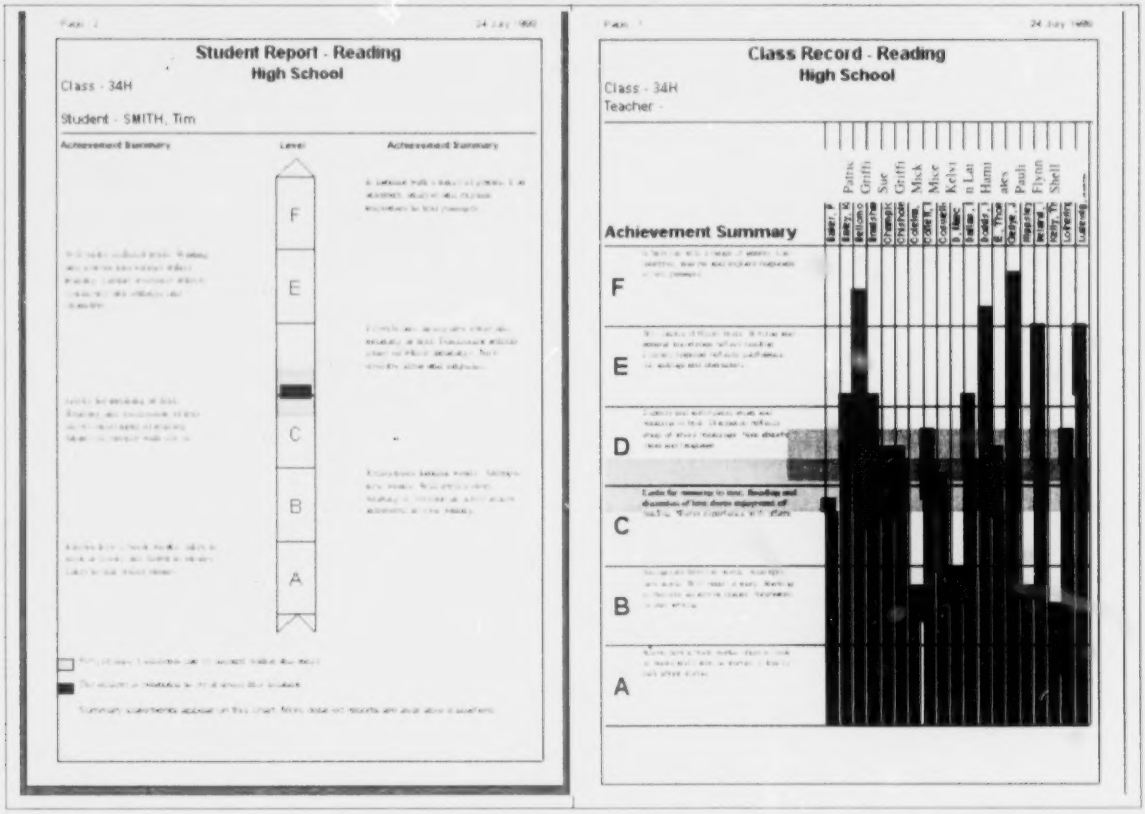
Interpretation involves identifying clusters of items or scores points on the right of the chart. A content analysis of the scoring rubrics and task content within each segment yields a description of progress against outcomes that define the competency of Working Numerately. These descriptions become the levels of increasing competence we use to interpret the students' work. There is no need for a numerical score in the report.

BEST COPY AVAILABLE

Descriptions form performance profiles. Each one in the example above consists of two parts: the first describes the outcomes and skills involved in the student's performance, and the second illustrates how the skills are used. With the use of computer software (Griffin 1997) that allows schools to generate reports, the descriptors then become the basic information for reports to a range of audiences, including the government, the school, the teacher, the student and parents. The school-based reports are illustrated in Appendix A.

The Student's Report

This report is for the student's portfolio. It needs to be accompanied by samples of work that illustrate the level of progress. It combines the descriptors and the box (from the box-and-whisker plot), and adds the student's level of progress. The rocket (*see* Appendix A) was an innovation of my colleagues working on the literacy profiles project in 1987. It combines several ideas. Teachers working with students and parents with non-English-speaking backgrounds were frustrated in attempts to describe how students were progressing—all reports and descriptors were written in English and described performances in the jargon of the classroom. Teachers simply drew a line, marked on it the letters A, B, C, etc. to represent the levels and then, with a felt pen, drew a region that roughly indicated the level to which the student had progressed, adding another line to show where the class was in general. My colleagues took this idea and combined it with what we were calling the "nutshell" descriptors of the profiles, based on the charts developed by Peter Hill and his colleagues for the TORCH test (Mossensen, Hill and Masters (1987) that we were using to scale teachers' judgements on the literacy profiles. The rocket reports directly what students can do, and how they compare with a peer group. The descriptions can be translated to cater for non-English backgrounds. More than a single report is required.



The Parent's Report

This report displays the five or six strands of numeracy (AEC 1995)—or any other area of learning—and uses the box-and-whisker plot again. The descriptor for the student's level of progress for each strand is printed, and we have allowed a space for the teacher to add a comment in the student's computer record. All these are written in an overall subject profile report that is sent to parents.

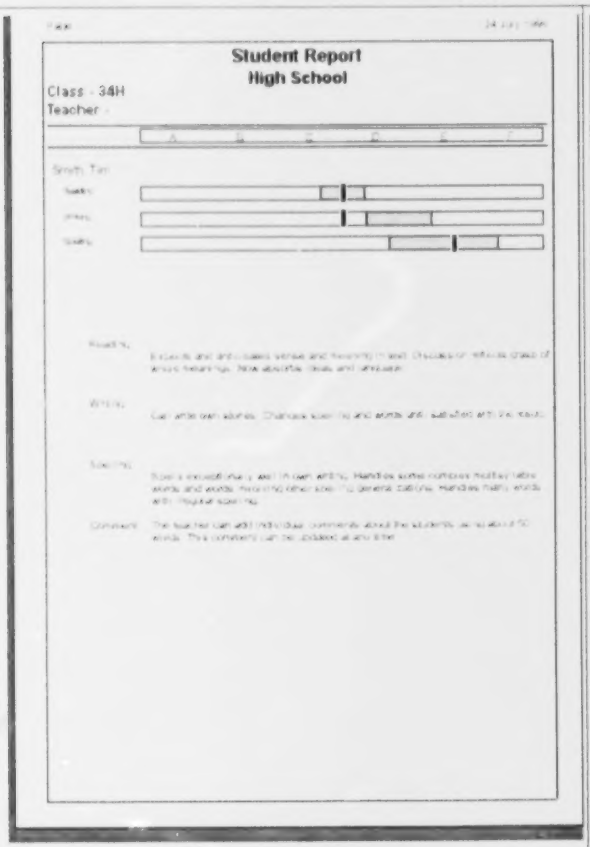
The Class Teacher's Report

The teacher can generate a report that shows, for each strand, every student in the class. The box is drawn across the chart so that the teacher can interpret what each student can do, how he or she compares with a peer group (which could be the class, the state, or both) and how students compare with each other. Using a second colour (the caps on the columns and the smaller and higher box), we add another set of records for a second assessment to be obtained at a later time. The chart now illustrates all the achievement information plus the relative rate of progress. The overlap between the two boxes illustrates for the teacher the average development for the class group. Armed with this information, the teacher is better prepared to intervene in the case of specific students or groups of students, or even in the teaching strategy.

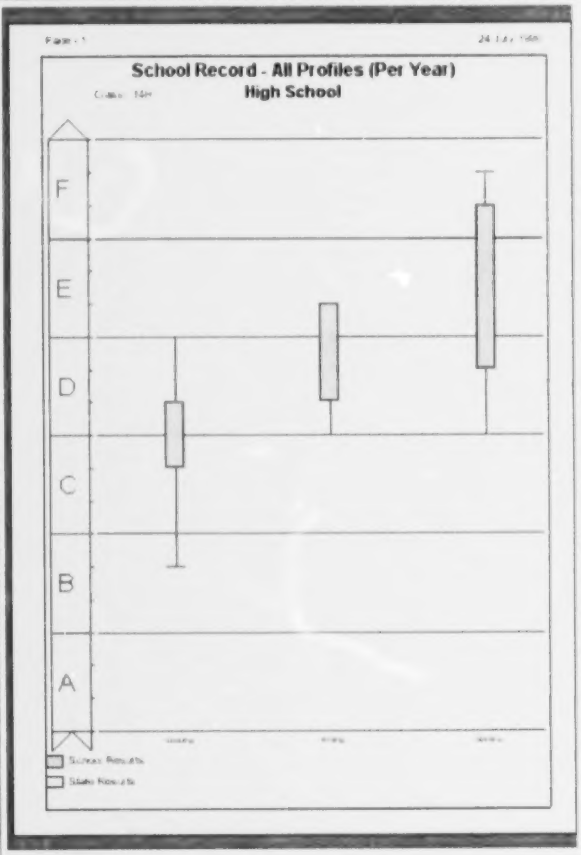
BEST COPY AVAILABLE

The School Administrator's Report

School administrators need to present reports on student progress without detail on individual students. The report and analysis of the performance data now presents them with an overview of subjects by grade, compared with state level (or any other group) distributions. At this level, curriculum decisions can be made. What we know about this type of profile reporting and record keeping is that they lead to changes in teacher discussion about student progress, and to more specific observation of progress in class (Smith 1996).



Parent Report



School Administrator's Report

A report needs to take account of several things: the author needs to be aware of the audience, their information needs, consequences of the communication and the effect of the medium of the report.

For the workplace assessment there is a different set of audiences, each with their own range of information needs; they might include the employee, the supervisor, the employer, the industry, and state or national bodies. Several interesting developments in

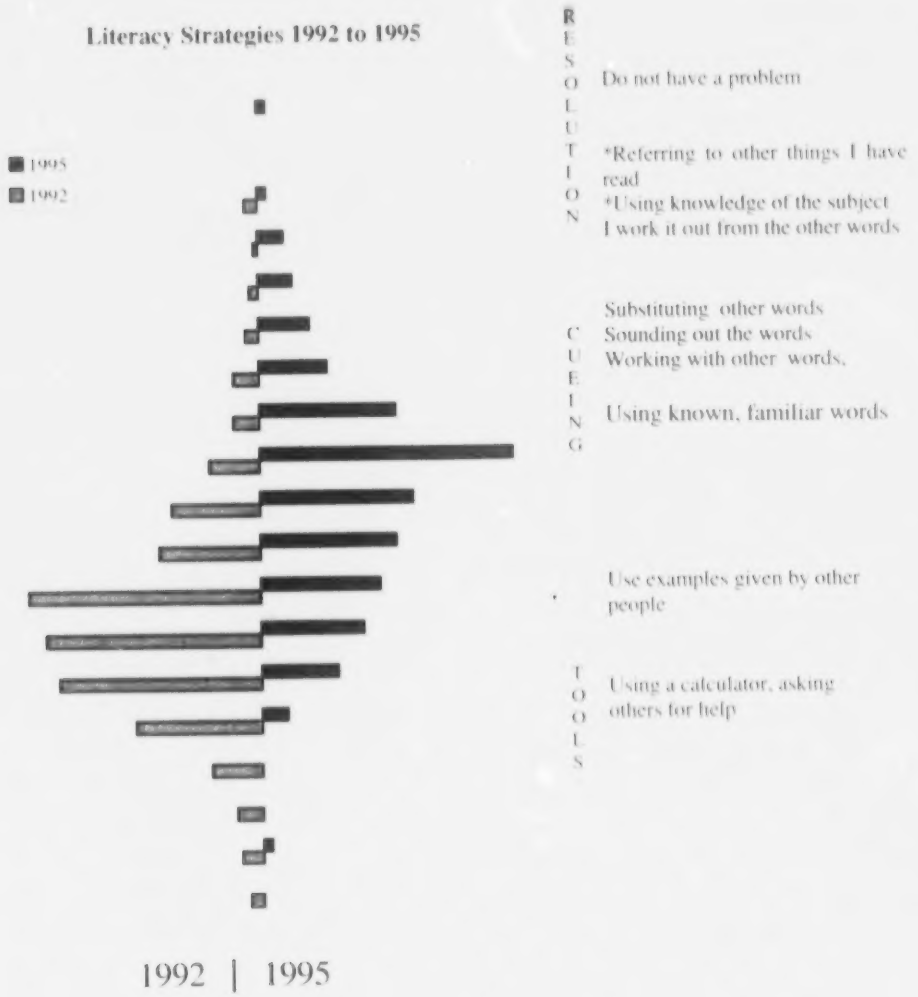
BEST COPY AVAILABLE

workplace assessment reports are taking place with the introduction in Australia of a skills passport in either paper or electronic format (smart card in Singapore) which lists the competencies acquired. In another development, associated with the on-line delivery of training, assessment and reports are being administered and analysed via the internet (Griffin 1996b). We will deal more with this in a later section.

Assessment and Judgement

Not every assessment requires a test, or even a separate observer. Self-assessment can be a powerful method of ascertaining performance. The literature on self-assessment is consistent in its evidence that, where stakes are low and include the opportunity for verification, assessors tend to under-estimate their own performance in comparison with assessments using external procedures. However, self-assessment is not likely to be seriously entertained in high-stakes assessment. The political and administrative fascination with instruments, because of their standardised application and objective scoring, has persisted for a century. There is no doubt, however, that teachers believe they are superior judges of learning outcomes and appropriate strategies. The use of the literacy profiles and the work by Hill, Rowe and others for almost a decade has shown that judgement using a holistic frame of reference—such as that provided by profiles—leads eventually to high levels of reliability and a level that can equal and even exceed that of some standardised tests (Rowe and Hill, 1995).

Self-assessment can be developed in a number of forms. Another example combining an instrument with self-assessment has been demonstrated in our work in the assessment of adult literacy, in which participants were interviewed in 1992 and again in 1995 (Griffin, Pollock, Corneille and Fitzpatrick 1997). The interviews sought information about the strategies the adults used in resolving literacy and numeracy problems in everyday life. When the data were scaled using item-response modelling, and then equated over the two periods, an exciting outcome emerged. On each occasion, the pattern of strategies was consistent in moving from referring to others, using support tools and standard cueing systems, to independence. It also became clear that a shift in the distribution of strategies had occurred, yielding evidence that the sample in general had learned to read.


Literacy Strategies 1992 to 1995
1995
1992
RESOLUTION
Do not have a problem
*Referring to other things I have read
*Using knowledge of the subject I work it out from the other words
CUEING
Substituting other words
Sounding out the words
Working with other words,
Using known, familiar words
Use examples given by other people
TOOLS
Using a calculator, asking others for help
1992 | 1995

In 1992, very few of the participants were aligned with the capacities to use cueing systems. Most relied on getting assistance for basic information. By 1995, the alignment of individuals was such that most were using cueing systems. This process gave evidence that the participants were using reading strategies, and by inference, had differences in relative reading abilities. Without testing in comprehension, it allowed us to monitor movement along a continuum, using levels of increasing competence to interpret the respondents' self- assessments. Self-assessments need not be based solely on judgements of ability. They can focus on process.

Using multilevel modelling (Rabash and Woodhouse 1995), another measurement model now becoming increasingly common in assessment research, we saw that 16 percent of variance in the 1995 strategy use was explained by group membership alone. The 1992 reading strategy explained a further 3 percent of the variation. This was four years after the initial program. It indicated that, while the initial strategies were important, the group, the program and/or the provider that offered the adult literacy program was important as a predictor of later development of reading ability. It was as if

the adult literacy programs were *preparing participants to learn* to read rather than *teaching them* to read. More important, it appears that it mattered a great deal which program was attended, because some helped to prepare the adults as future readers much more than others. There is no doubt that value added analysis such as this is a way of identifying the programs and characteristics of programs that are related to improved achievement. Reducing analyses to simplistic comparisons of means, medians, or box and whisker plots is not an optimal means of identifying ways to improve.

Instruction

Performance assessments associated with outcomes in schools need standards to which students and teachers can direct their efforts. They must maintain their validity, especially when they dominate instruction and curriculum. And they must be capable of exemplifying standards by setting clear targets for instruction and learning efforts. Deconstructing knowledge into disconnected, decontextualised bits makes it difficult to assess capabilities. The application of psychometrics and of item response models is leading to better understanding of levels of learning and to the direct interpretation of performance. The dilemma between contextual fidelity and the need to generalise seems to be predominantly resolved in the favour of the context. The issue of context, however, is a new area for research; instead of treating contextual influence as "noise", Resnick and Nolan (1995) are leading developments in assessment that allow for both contextualisation and generalisation.

Most of the negative aspects of assessment can be obviated if the role of prediction in assessment is recognised and insisted upon. If the test, the performance assessment or the portfolio cannot help to predict meaningful performances to identify learning needs in context and to inform teaching and learning, the consequential nature of the assessment must be questioned. The meaning of a score, an average, a range—or, indeed, a qualitative description—that does not address the knowledge, skills and attitudes exhibited, and the way these forecast future or generalised learning or performance, should be challenged. The problem may not always be the assessment, of course; the outcome statement itself may be deficient. If these basic building blocks are neither clear nor meaningful, the assessment may not be worth doing.

Clearly, outcomes-based education involves radical changes to the ways teaching and learning operate. The shifts away from a time-based orientation to a "continue till succeed" approach means that the monitoring of outcomes is fundamental to each of the outcomes orientation in education. Without continuous monitoring, neither can succeed. When viewed in this light, outcomes need to be important and generalised, and to lead to changes in the learner that can, through assessment, be generalised in much the same way as was discussed at the beginning of this paper.

Outcomes-based education and assessment demand that individuals are able to show that they have learned. They focus on demonstrable changes in the learner, and not on transmission of content. Opponents argue that they are behaviourist in approach, and that, when all the outcomes have been achieved, some intangible aspects of human aesthetics are missing. The argument is that there are some parts of learning that cannot be observed or measured; that assessment reduces education to the level of direct

observation and measurement; moreover, that incidental learning, so valued by teachers, is lost, and that the assessment of student performance becomes the central issue of education. This is an impoverished view of learning, because performances are a means through which learning is inferred (Willis & Kissane 1997). By identifying the cohesive sets of tasks—rather than isolated single tasks—we are able to plan far more efficient development.

References

Adams, R., & Khoo, S.T. (1995). *Quest: an interactive item analysis program.* Melbourne: Australian Council for Educational Research.

Biggs, J.B., & Collis, K.F. (1989). Towards a model of school-based curriculum development and assessment using the SOLO taxonomy. *Australian Journal of Education, 33*, 151–163.

Bloom, B. S. (ed.) (1956). *The taxonomy of educational objectives: The classification of educational goals, handbook 1: Cognitive domain.* New York: Longman.

Board of Studies (1995). *The curriculum and standards framework.* Melbourne: Author.

Browne, B. (1997). *Assessment driven instruction.* Paper presented at the Assessment Research Centre Seminar series, University of Melbourne, 20 June.

Eisner, E. (1993). Why standards may not improve schools. *Educational Leadership, 50*(5), 22-23.

Gangé, R. M. (1970). *The conditions of learning* (2nd ed.). New York: Holt.

Gillis, S., Griffin, P., Trembath, R., & Ling, P. (1997). *The theoretical underpinnings of assessment* (Research Grant Report). Melbourne: Assessment Research Centre, University of Melbourne.

Gillis, S., Griffin, P., & McPherson, C. (1996). *A review of the national skills recognition system for the transport and warehousing industry.* Melbourne: Assessment Research Centre, RMIT.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist, 36*, 923–936.

Griffin, P. (1990). Profiling literacy development: Monitoring the accumulation of reading skills. *Australian Journal of Education, 34*,, 290–311.

Griffin, P., & Smith, P. (1996a). *The impact of outcomes-based education on teachers' work.* Melbourne: Assessment Research Centre, The University of Melbourne.

Griffin, P. (1996b). *An interactive on line assessment system for outcomes based assessment for special educators.* Proceeding of the Australian Resource Educators Association Annual Conference, Ormond College, University of Melbourne, September 11–12.

Griffin, P. (1997). *The ARC learning profiles (Version 1.1) [Windows 3.1 and up.]* Melbourne: The Assessment Research Centre, University of Melbourne.

Griffin, P., & Jones, C. (1987). *Assessing the development of reading behaviours: A report of profiles and reading band development*. Paper presented at the Annual Conference of the Australian Association for Research in Education, University of New England, Armidale, NSW, November.

Griffin, P., & Nix, P. (1991). *Educational assessment and reporting: A new approach*. Sydney: Harcourt Brace Jovanovitch.

Griffin, P., Pollock, J., Corneille, K., & Fitzpatrick, M. (1997). *Skilling me softly: A study of the destinations of adult literacy students*. Melbourne: Assessment Research Centre, University of Melbourne.

Harste, J. (1992). *Changing practices in assessment*. Newark, Delaware: International Reading Association.

Jasa, S., & Enger, L. (1994). Applying OB to arts education. *Educational Leadership, 51*(6), 30–32.

Marsh, C. (1995). The tasks and objectives of teaching. In R. Dullahide (Ed.), *Educational psychology: An Australian and New Zealand perspective* (pp. 28–44). Brisbane, Australia: John Wiley & Sons.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

National Training Board (1992). *National competency standards*. Canberra: National Training Board.

Office for Educational Review (1997). *Year 9 numeracy assessment monitoring program: Task 4*. Hobart: Office for Educational Review, Department of Education, Community and Cultural Development.

Popham, J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, N.J.: Prentice Hall.

Rabash, J., & Woodhouse, G. (1995). *MLn command reference*. London: Institute of Education, University of London.

Rasch, G. (1980). *Some probabilistic models for the measurement of attainment and intelligence*. Chicago: MESA Press.

Resnick, L., & Nolan, K. (1995). Where in the world are class standards? *Educational Leadership, 52*(6), 6–10.

Rowe, K., & Hill, P. (1995). *Assessing, recording and reporting students' educational progress: The case for subject profiles*. Paper presented at the Annual Conference of the Australian Association for Research in Education, Newcastle, Australia, November.

Rowley, G. (1982a). Historical antecedents of the standards setting debate: An inside account of the minimal beardedness controversy. *Journal of Educational Measurement, 19*(2), 87–95.

Rowley, G. (1982b). Response to Livingston's comment on minimal beardedness. *Journal of Educational Measurement, 19*(3), 231–33.

Sadler, R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*(2), 191–209.

Slavin, R.E. (1994). Outcomes-based education is not mastery learning. *Educational Leadership, 51*(6), 14.

Smith, P. (1996). *Self regulating meta-cognition of readers in grades 5 and 6.* Unpublished Ph.D, RMIT, Melbourne.

Vygotsky, L.S. (1962). *Thought and language*. Cambridge, Mass.: Massachusetts Institute of Technology Press.

Willis, S., & Kissane, B. (1997). *Achieving outcome-based education: Premises, principles and implications for curriculum and assessment.* Perth: School of Education, Murdoch University.

COMMENTARY ON PATRICK GRIFFIN'S PAPER

Cedric Croft

New Zealand Council for Educational Research

In drafting a commentary **before** I had the opportunity to read Patrick's paper, I began in this vein:

"It is clear from Patrick and elsewhere, that there are common threads running through the assessment policies and debates within countries that we as New Zealanders identify with quite readily. Here I refer to Australia, Canada, the U.K. and the U.S., all of which have been said by the current Minister of Education, to have approaches to national assessment that we should follow. These common threads include:

- a perception that schools 'could do better' on behalf of some identifiable groups of students;
- public concerns about standards of literacy and numeracy, without necessarily defining what these standards are;
- a wish to make schools more accountable for their stewardship of public funds;
- a *concern* that schools do not merely mirror many of the negative attributes seen in society at large, and a *belief* that declines in standards begin in schools.

How then are these tendencies to be overcome?

- by determining curricula centrally and representing 'school effectiveness' in terms of the achievement of students;
- by instituting external testing to identify groups at risk, and providing a common yardstick for use *within* and *between* schools.
- by broadening previously-established systems of national or statewide testing to include performance assessment, in an attempt to assess more validly the range of skills actually taught in many classrooms;
- by promoting more effective formative assessment (also known as Assessment for Better Learning);
- or by some particular combination of all these elements, which seem to me to be exactly the recipe being stirred up in New Zealand.

The New Zealand Cooking Pot?

How will these ingredients mix in the New Zealand cooking pot? Or is it the New Zealand melting pot? Or perhaps these days, it might be a wok!

We have a particular curriculum structure and a particular set of national priorities and beliefs, all wrapped in a particular set of assessment priorities. So 'recipes' from overseas need to be looked at closely before they are adopted.

Will the recipe work?

- External testing of the type put forward in the Green Paper of 1998 seems not to have worked overseas. Why should it work here? New Zealand is promoting a broader mix incorporating diagnostic testing, development of exemplars of student performance and examples for teacher assessment, some of which will be incorporated within the

ARBs, plus additions to the current NEMPs. The mix here is much broader than straight national tests. So the mix might work, but which parts will be given credit?

- A centralised and controlled curriculum backed up with very specific curriculum testing would work (if the test scores were to be believed!), but are the children receiving an education? Who might engage with a curriculum like this? Who would teach from this document?
- Performance assessment *is* the staple approach within compulsory schooling at present, and this has not had to be introduced as part of a mandatory national system. A major strength of school assessment in New Zealand is that it does duplicate well what children are doing in most classrooms."

But clearly, Patrick's paper demands a different response, even if what I have noted is believable and based on international trends. He reminds us of:

- the primacy of *validity* and that it is *inferred* not measured directly;
- that reliability (this may be measured) is a prerequisite for validity;
- that the breadth of current curricula and the drive to raise standards universally has outstripped the utility of standard psychometric techniques in many instances;
- that in an assessment context there are only *four* things we can ask students to do. These are:
  * write
  * make
  * do or
  * say something

  and that at least *three* of these pose some almost intractable assessment difficulties;
- that we have three frames of reference for *interpreting* assessment:
  * norm
  * criterion
  * ipsative
- that reporting is the crucial follow-on from assessment, as it is by reporting that results from assessment may be turned into positive action;
- that the primary aim of assessment is to *predict* something, and our predictions are based on *inferences*;
- and that *above all* we must go into any assessment activity, whether it be national testing, school-wide assessment or single classroom assessment, with clear purposes in mind.

His points about criterion referencing are timely, as there is much about this general construct that is widely misunderstood. Levels of performance, or where cut-scores are set for mastery tests, are *not* the essence of criterion-referenced assessment—the essence is assessment that is interpreted through the task(s) performed. Glaser (1987) later refined this as, "the development of procedures whereby assessment of proficiency could be referred to stages along progressions of increasing competence".

Key questions such as the list that follows can be useful in helping to clarify assessment purposes:

Key* Questions to shape Assessment Plans

WHAT

- are we setting out to do?
- * are we hoping to achieve?
- resources are we going to commit?
- else is competing for these resources?
- expertise is available to us?
- * do we hope to alter as a result of this assessment?

WHY

- * are we undertaking this assessment anyway?
- will we do it in this particular way?

HOW

- will we go about this assessment?
- * will we use the information we gain?
- will we use the expertise we have?
- will we find complementary expertise?

WHEN

- will this assessment be undertaken?

WHO

- is to be involved?
- * needs to know the outcome?

A PRIMARY PERSPECTIVE

Frances Nelson

Principal, Northcote Primary School
NZEI National Executive

Introduction

Thank you for the opportunity to contribute to the debate with researchers, policy makers, and pre-service educators, so that with the teaching profession we can move forward jointly on this crucial issue. My brief is to present a primary perspective and I am pleased to be able to do that from my current position as Principal of Northcote Primary School and as an NZEI national executive member.

We are all passionately interested in this topic because we recognise that assessment procedures are a critical component of the education process, and assessment is one of the most powerful educational tools for promoting effective learning, if approached in the right way.

Assessment for Learning

Paul Black and Dylan William (1998) conducted a review of research on the impact of classroom assessment on learning. The review showed that, when carried out effectively, informal classroom assessment with constructive feedback to the student will raise levels of attainment. Although it is now widely accepted that this form of assessment and feedback is important, the development of practice in this area will need to be supported by a change in policy that identifies and supports this as an appropriate model. The push for better learning is evident in aspects of current government policy.

Recently the Assessment Reform Group in Britain followed up the work of Black and William and took it further. The Group identified the problem that much current practice emphasises assessment *of* learning rather than assessment *for* learning, and misses the opportunities to use assessment to improve learning. The research indicates that improving learning through assessment depends on five deceptively simple key factors:

- the provision of effective feedback to students;
- the active involvement of students in their own learning;
- adjusting teaching to take account of the results of assessment;
- a recognition of the profound influence assessment has on the motivation and self-esteem of students, both of which are crucial influences on learning;
- the need for students to be able to assess themselves and understand how to improve (Assessment Reform Group 1999:4–5).

Primary schools are working in the mode of *assessment for* learning. In the context of rapid curriculum change, there is work to be done to effectively integrate the assessment process into curriculum implementation. This work is ongoing and the task is to make these factors consistent and used effectively. The range of different approaches to the design of the curriculum documents means that this will take time to consolidate.

Teachers can be expected to make changes only if they are encouraged to do so within a supportive environment. The demands made of them by management and government therefore need to facilitate these changes, not constrain them. Real change in education occurs only when there is commitment by classroom practitioners. Experience tells us that this commitment will be given when teachers understand the work required and see that it is educationally sound and relevant to their work. Government agencies can do

much to support changes in classroom practices through supporting the cooperative culture schools look to develop, and by providing the resources to do the job.

The Education Review Office (ERO) has had a powerful influence on practice in schools. As a result of self-management, there has been an increased emphasis on schools collecting information to show that students are reaching desired levels of achievement. The surveillance of schools by ERO and the possibility that the results of such reviews will be published in the media has resulted in assessment becoming a high stakes activity. A great deal of anxiety about this type of assessment exists among teachers.

In its report on *Student Assessment: Practices in Primary Schools* (1999), ERO found that classroom teachers currently spend a great deal of time and effort on assessment. But it also claimed that, in the main, this assessment is not of the quality needed to help teachers determine individual student learning needs, progress and achievement, or to give accurate information.

Hill (1997) provides more detail of what is going on in primary schools. She argues that rather than summarising information collected during formative assessment to give a broadly based but brief portrait of the student's achievement, many schools have developed detailed objective by objective profiles about individual students which duplicate the information already captured. The main reason this was done was a genuine attempt to meet the needs of people other than the teacher and the student. Teachers voice a commonly held view that summative information is mostly for quality control or evaluative purposes, rather than being a way to summarise information about particular students at particular times.

ERO must make clear what it expects schools to be doing in relation to this aspect of their practice, and how it will view innovative procedures that are being tried by schools. Alternatively, schools must be audited against their own assessment objectives and practices as self-managing schools. A level of consistency must prevail, so that schools are responding to the needs of their school population and community, rather than to compliance issues. This will enable a focus on *assessment for* as opposed *to assessment of* learning, as described by the Assessment Reform Group in Britain.

Meanwhile, for teachers there is an increasing frustration that all this diverts time away from teaching.

National Testing Proposal

In the light of this, the current government's continued focus on introducing annual national paper and pencil tests is disappointing. It would be unprofessional not to challenge this aspect of government's policy.

We know all children operate at different developmental levels in different skill areas throughout the year. Rather than being narrowly compared to each other, as a National Test would do, learners should be observed, coached and encouraged in their growth. Some parents recognise the trends in their child's learning, but the damage is much more personal and serious when too much is attached to one score on a narrow range of skills. For many parents, their child's score will not measure just one day or even a week of their child's work. It mirrors the family success and the child's future. And how should they

know otherwise? Most parents' sole form of knowledge of tests of this type springs from their own first-hand experiences with the examination process at secondary school level.

Over time we have learned from the pattern overseas that the tests become institutionalised; the scores drive many school initiatives and curriculum decisions. What is worse, the testing regime inevitably expands as people become resigned to its presence.

It is our strong belief that it would be more fruitful to find a common language for parents and teachers to talk together, and to provide teachers with assessment tools which will encourage and enable this discussion. This is how true partnership in children's learning will develop.

To make this type of process work, a comprehensive professional development programme will be needed to ensure classroom teachers have the skills and understanding to carry it out. This would need to be supported by a parent/community education programme to enable full participation from all of the partners involved in the education of children.

An intensive, interactive assessment process of this nature will require resources to enable its success. Workload of teachers is a critical issue. Teachers are struggling to carry out the required activity to successfully implement the type of teaching/learning/assessment work that we are describing. In schools throughout the country, teachers are replicating each other's work without being able to use their knowledge collectively with other schools in an organised way.

To make this work, class sizes and workload must be addressed. Without attention to this, any system reflecting the qualities outlined above will be doomed to failure. The government's commitment to filling the "information gap" for parents and teachers will be measured by their preparedness to take the route which will bring educational success rather than following the easiest or cheapest path, i.e., via a National Test.

Ways Forward

NZEI argues that the focus needs to be on helping teachers to use assessment as part of teaching and learning, and in ways that will raise student achievement.

The Assessment Reform Group (1999:7) summarised the characteristics of assessment that promote learning:

- it is embedded in a view of teaching and learning where it is an essential part;
- it involves sharing learning goals with students;
- it aims to help students to know and to recognise the standards they are aiming for;
- it involves students in self-assessment;
- it provides feedback which leads to students recognising their next steps and how to take them;
- it is underpinned by confidence that every student can improve;
- it involves both teachers and students reviewing and reflecting on assessment data.

We need to build on these findings, not invent a crisis that doesn't exist. Neither should we accept or support an imposed system of national paper and pencil tests which does not lead to better teaching and improved learning.

We know what is required. We have a clearly identified range of options to provide for our needs. We have had the debate nationally and examined the research and experiences internationally. We now need to be able to get on with it.

References

Assessment Reform Group (1999). *Assessment for learning. Beyond the black box.* Cambridge:University of Cambridge School of Education.

Black, P. & William, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1).

Education Review Office (1999). *Student Assessment: practices in primary schools.* No. 2, Winter.

Hill, M. (1997). Planning valid and valuable school assessment systems: a possible model for New Zealand schools. *Teachers and Curriculum, 1.*

Wylie, C. (1997). *Self managing schools seven years on. What have we learnt?* Wellington: New Zealand Council for Educational Research.

ASSESSMENT FOR THE SECONDARY SECTOR

Graeme Macann

PPTA President

An overwhelming priority for the secondary sector is for us—the practitioners, and the Ministry and the Qualifications Authority especially—to develop a consensus around an assessment and qualifications policy for the senior part of our secondary schools. It is unfortunate that the territory is so poisoned now by the mistakes, missed opportunities and failures of the last decade. Attributing blame for those can be superficially attractive and diverting, but it seems to me to be particularly unhelpful because it results in simplistic rhetoric being hurled from one end of the sector to the other. I have to confess to having participated in that practice on occasion. It is time to move more constructively to solutions.

The results of missed opportunities, stuff-ups, call them what you will, cannot be ignored, however. One result I cannot ignore is that a number of teachers feel cynical and disempowered with regard to assessment and qualifications reforms, often for quite different reasons. For example, some of those who were innovative with unit standards now feel betrayed that their work will, to some extent, have to be undone if Achievement 2001 and the new National Certificate of Educational Achievement (NCEA) come into being. Recent history seems to have taught those who mostly ignored the Qualifications Framework that doing so was a sensible strategy. Helping to move both of those groups, and others, to an active, alert and lively engagement with the issues is a difficult goal, but one that is essential to strive for.

I want to discuss briefly some criteria for evaluating qualifications systems, and use them with reference to the proposed NCEA. The criteria I will use were published in *Te Tiro Hou*, a report of the Qualifications Framework Inquiry, in June 1997. This report was commissioned by PPTA, and the criteria that its authors (Peter Allen, Terry Crooks, Shona Hearn, and Kathy Irwin) developed became part of PPTA policy later in 1997.

The QFI report argued that an effective assessment and qualifications system would be:

- Fair
- Inclusive
- Cumulative
- Clear
- Motivating
- Coherent
- Constructive
- Manageable. (Allen, Crooks, Hearn and Irwin 1997:95–100)

1. **Fair.** This implies that the credits and qualifications accurately describe learner achievement, and are trusted. Learners should be given a fair chance of demonstrating their achievement. Standards should be sufficiently consistent to ensure all learners are awarded similar results for similar levels of achievement. Finally, how student achievement is described should communicate as clearly as is reasonably possible the standards achieved.

Will the NCEA be fair?

The overall design feature of expressing what will be assessed in terms of standards implies more fairness than the current norm-referenced School Certificate and Bursary Examinations.

We do not yet know the details of the moderation system for the internally assessed achievement standards, so that aspect of fairness is an unknown. There will be more detailed description of what students have achieved, via credits and levels of credit, thus more will be communicated to those who need to interpret that information.

2. **Inclusive.** "Credible pathways are needed to worthwhile qualifications for as many learners as possible" (*Te Tiro Hou*, p. 96). Growing numbers of students are undertaking a sixth and seventh form year. This highlights the deficiencies of current awards, and the fact that many alternative courses offered by schools do not contribute to a coherent qualification.

   Will the NCEA be inclusive?

   The NCEA should be a major advance on what we have at the moment. Students will be able to be assessed via a variety of achievement and unit standards, and accumulate credits from a multiplicity of courses leading to one qualification. The Leaders Forum[1] proposed that the widest possible range of achievement and unit standards be available for credit accumulation.

   The NCEA being inclusive also means that potentially the sector will mostly rid itself of the sins of having high status and low status qualifications in competition at the same level.

3. **Cumulative.** At the moment we have three discrete qualifications—School Certificate, Sixth Form Certificate and Bursary—in the senior secondary school. There is no common currency between them, and they cannot be linked to other qualifications that senior school students are commonly studying for.

   Will credits achieved in one context be accepted in others? In other words, will the NCEA be cumulative? Clearly this is a goal of the new system, and one that has won widespread acceptance.

4. **Clear.** Students and teachers need to understand as clearly as is possible the knowledge, skills and standards expected. Tight specification of outcomes may not be possible, and ought actually to be avoided if achieving them means narrowing the goals of a course in a way that dilutes the challenges.

   Will the knowledge, skills and standards expected in the NCEA be clear?

   There is more information in the draft achievement standards than exists in the current examination prescriptions. Critics may argue that there are no clear standards in the achievement standards. However, the promise of exemplars, at least one pre-moderated assessment task per achievement standard, and professional development should help to make the standards clear. It is essential that the Qualifications

[1] A group convened by the Ministry of Education, consisting of secondary principals, members of PPTA's Board of Studies and others. They met on 7–8 September 1999, and were due to met again on 1–2 November 1999.

Development Group are given the resources to do an excellent job with the above. It is worth noting that we expect excellence from students and teachers; let us demand it from the policy developers as well.

By comparison, at the moment there are no exemplars for standards in School Certificate, Sixth Form Certificate and Bursary. Guinea pig scripts used by previous marking panels for the national examinations would be an excellent resource, but these have never been distributed. Similarly, a marking schedule does provide considerable guidance, but understandably these are made available only after students have sat the examinations and their contents are not usually shared with future students, at Form Five level at least.

5. **Motivating.** "Learners are much more likely to put continued effort into their learning if they can see pathways ahead which make sense to them and seem worth following" (Allen, Crooks, Hearn and Irwin 1997:98). As well, learners need to sense a genuine prospect of sufficient success to justify a sustained effort. Students have to be challenged, but the targets that challenge them have to be attainable. As well, excellence has to be rewarded and celebrated.

   Teachers need to find it motivating as well. This implies some scope for professional freedom, that assessment will not dominate the curriculum, and that there will be appropriate professional development and support materials.

   Will Achievement 2001 be motivating?
   We do not know yet, because some decisions that are crucial to that aspect have not yet been made. For example, PPTA believes that the entry level for Level One credit must be set low enough to allow a large proportion of students in Form Five to earn credit for most achievement standards. Expectations for merit and excellence will, of course, be more demanding.

   Similarly, the nature of the moderation system for the internally assessed achievement standards is not yet fixed. If it proves as cumbersome and unresponsive as the current unit standards system often is, then teachers will find it very unmotivating.

   Excellence has to be rewarded, and the prospects for that so far are promising, both in terms of excellence being achievable in each achievement standard, and also with the possibility of one scholarship system prevailing, instead of the duality of the state and private systems we have at present.

6. **Coherent.** Everyone, teachers, students and parents, needs to be able to see the "big picture" and to see linkages. There is a temptation that easily measured outcomes will dominate at the expense of more complex outcomes, which are more difficult to assess.

   Will the NCEA promote coherence or fragmentation of learning and teaching?
   If we can get through our current obsession with assessment and refocus on learning and teaching, there is every chance that the standards will be used in a way that is coherent rather than atomised.

7. **Constructive.** A good qualifications system should promote effective learning. Learners and teachers need clear and usable feedback on progress. Ideally learners will be given more than one chance to demonstrate that they can attain the required standards.

   Will students (and parents, and other users of qualification information) receive clear and useful feedback on progress with Achievement 2001?
   There are concerns at the moment about the potential for raw-score, percentage or percentile recording to be confusing to those who need to understand what is being reported, and for internally and externally assessed standards to be treated differently. Putting those issues to one side, reporting by achievement or unit standard is capable of providing more useful feedback than occurs at present.

8. **Manageable.** The assessment procedures need to involve acceptable levels of stress for students and teachers within acceptable levels of expenditure by the community.

   Will the whole assessment regime under Achievement 2001 be manageable?
   This is a huge issue in terms of the quantity and nature of the assessment events that students would face, and for teachers in terms of their own workloads. The ratios of internal and external assessment for the achievement standards will reflect concerns about, and impact upon, the teacher workload issue to some extent.

   Having a significant proportion of external assessment at Sixth Form level ought to reduce teacher workloads. How that will be balanced against more high stakes internal assessment at Fifth and Seventh Form levels is an open question at the moment.

   The Ministry providing a good quality assessment bank will be important, as will be a commitment from them and Boards of Trustees and the practitioners to excellent initial and ongoing professional development.

Conclusion

Will the NCEA be a horse designed by a committee that looks like some horror created by Larsen? Is it a workable compromise, or will we end up with the worst of the exam and unit standards systems?

The jury of our members is still listening to the case. At our Conference in September there was considerable debate on this topic. We were able to report a good deal of dialogue with Ministry and Qualifications Development Group officials, and a necessary degree of openness on their part to suggestions for changes to the NCEA. The Ministry and PPTA have since jointly developed and run a consultation exercise on NCEA proposals with representatives from all secondary schools.

We have been given some rope to continue working on the proposal. But there is a choker on the end of the rope!

Reference

Allen, P., Crooks, T., Hearn, S., & Irwin, K. (1997). *Te Tiro Hou*. Report of the Qualifications Framework Inquiry. Wellington: PPTA, June.

ASSESSMENT—MAORI ATTITUDES TO ASSESSMENT

Wharehuia Hemara

New Zealand Council for Educational Research

Kia ora koutou katoa

I would like to repeat a story I heard at a seminar earlier in the year.

A Story

The scene was set in the U.K. Inspectors were assessing the cognitive abilities of some young pupils. An illustration of a train on a railway-track was placed in front of them. In the foreground the track was broken. If the train and tracks were real, inevitable disaster would result. The children were asked what they would do under the circumstances. Most came up with strategies for warning the train-driver. One little boy said, he would *tell his mum.* Apparently the inspectors were disappointed by this response. But one persisted in her/his enquiry and asked why he would *tell his mum.* He replied *because my mum has never seen a train accident.*

Different Perspectives or Cross Cultural Encounters

Although this story is set in another country and deals with a young child who lived, operated and was being educated within his and his family's cultural context, his response was not understood until further enquiry had been made. Maori children and their whanau run across this situation in their everyday activities. The difference between the young British child and Maori children is that Maori children and their assessors are often further apart. Assessment of their academic abilities is configured in such a way that Maori world views and knowledge of the universe are cloaked by the mores and values of a society which has difficulty understanding or accounting for Maori/Pakeha differences. A Board of Trustees has recently articulated this issue:

> . . .as long as Maori have been educated in the New Zealand education system and have been deemed as failing in it, we have been assessed using tools that we may not consider appropriate or even important (BOT, 19XX:16).

Comparing Like with Like

Although there are strong arguments for tacking Maori academic performance to socioeconomic conditions, the counter argument, that Maori possess cultural and by association cognitive differences is, I believe, just as compelling. Although it is tempting to compare the Maori situation to groups of poor whites in Britain, Afro-Americans, and Portuguese gypsies, it is more appropriate to draw common themes from other indigenous minority communities from around the world.

Indigenous Concerns and Cultural Misappropriation

I am about to take a leap of faith here! I would say that the cultural baggage which Maori and other indigenous groups carry, is quite different from that carried by other disenfranchised or marginalised groups. At the same time as strength may be drawn from the deeply held belief that they possess strong and valid claims to sovereignty over particular resources, there is a sense that their communities and values are under threat. They sometimes believe that outsiders treat them like some sort of *pick-and-mix*, where what the dominant society believes are attractive aspects of the culture are included in a national cultural side-show, and those that are not are either ignored or side-lined. Often that *pick-and-mix* is taken from a frozen and sometimes fictitious 18th and 19th century culture. While the attitudes behind those choices are often misty and romantic, the same attitudes are fed by news of academic failure, disparities on a number of fronts and a sort of schadenfreude horror of the *Once Were Warriors* kind.

Subversion and Resistance

On the other hand, worldwide indigenous resistance to Western European teaching and learning practices, which could be seen in the context of resistance to foreign hegemony, has influenced some Maori points of view:

> The history of the Maori encounter with schooling reveals that the education system has played a significant role in the securing of Pakeha economic and political dominance during the nineteenth century and in maintaining that dominance throughout the twentieth century . . . (Simon 1990:135).

This resistance to what was described in the *Appendices to the Journal of the House of Representives*, 1856 as ". . . a divine blessing" may be considered by many Maori as an almost subliminal act of subversion and resistance.

Diversity

What I am saying here is that Maori engagement in education has been rocky, but Maori culture is still very much alive and diverse. Maori who are native English speakers, and who have never been to a marae or do not know their whakapapa, are just as much part of contemporary Maori culture as those who have been steeped in ancient traditions. How assessment practices respond to Pakeha/Maori differences and Maori diversity is crucial.

The Green Paper

The Green Paper, *Assessment for Success in Primary Schools*, states that:

> Primary schools need assessment information that will help them remove barriers to learning by . . . clearly identifying the effectiveness of a school in improving achievement for Maori and Pacific Islands students (p. 13).

In the recently released Ministry of Education report *Assessment for Success in Primary Schools: Report of the Submissions to the Green Paper* (MoE 199X), responses to the Green Paper have been harvested from Maori parents, teachers and Boards of Trustees, most of whom are involved in Maori medium immersion education. While this is a positive sign that Maori are being consulted, it is worrying that no specific provisions for Maori learners in mainstream English medium education have been made. The Green Paper contains statements relating to Maori education generally:

> . . . Throughout the consultations on the Maori Education Strategy, people have emphasised that good assessment and evaluation systems are vital to show what works, and what does not work, for Maori students. An important first step in addressing improvements is knowing how well Maori students are learning in school . . . (p. 8).

The Paper goes on to say that there are a number of constraints surrounding Maori immersion education, which prevent full student assessment, and

> . . . There is a need for additional assessment activities appropriate for use in Maori immersion schooling, especially activities related to Maori language development. In the process of developing assessment instruments, it is important that Maori be consulted and involved . . . (p. 9).

Considering that most Maori learners are in mainstream English medium educational environments, it is to be wondered if associated Maori communities have been targeted in the same way?*!#. In the *Report of the Submissions to the Green Paper*, a statement by a Deputy Principal, which has been integrated into others relating to immersion education, does highlight some concerns about mainstream education: ". . . We need to be looking at the reasons Maori children don't achieve in mainstream education . . ." (p. 15).

However, it seems that Maori opinions were specifically targeted only when they were linked to Maori language immersion education. Because of that particular focus, only one section of the Maori community have had the chance to articulate their concerns as Maori.

Standards—The Same or Different?

Even though the following findings represent a limited catchment, they do display diversity. In a sample of 1658 respondents to the question "Should students in Maori-immersion schooling be expected to achieve to the same or different standards as those in mainstream education? Why?", 1384, or 83.5 percent, responded by stating that they believed learners should be assessed by the same standards as those in English medium mainstream environments. The reasons given were:

- no reason *9.9 percent*
- to be meaningful and effective *15.5 percent*
- but should be in te reo Maori *6.8 percent*
- competition is the same for everyone *12.9 percent*
- because what is being taught in both language environments is the same *7.3 percent*
- a matter of equity *13.4 percent*

- general hopes and aspirations are similar although the means to achieving them will be different *17.7 percent*.

Alongside some of these opinions were statements such as:

> Maori-immersion schooling should not be any less demanding than mainstream education, but it might have different emphases (BOT, p. 15).

> It would be appropriate to develop whatever is sensible, non-demeaning and acceptable to all races and whatever does not disadvantage any sector.

However, 274 or 16.5 percent responded by suggesting that Maori immersion learners should be assessed by different standards. The reasons given were:

- no reason *0.5 percent*
- different cultures and values are involved and so Maori should not be expected to conform to Pakeha standards *10.7 percent*
- tests are culturally biased and therefore do not reflect the true abilities of diverse groups of learners *4.2 percent*
- standards should be related to individuals and not be comparative *1.1 percent*

A statement from the Academic Board of the Dunedin College of Education supports some of the above:

> The nature of immersion schools implies different pedagogy, context, expectations and values. Assessment standards must, therefore, be different from those in mainstream education . . . Assessment standards must . . . be relevant to the curriculum, appropriate to the student and consistent with cultural values . . . (p .16).

Final Word

I would like to round off this short presentation with an excerpt from a letter from John Ormsby (Bob Mahuta's tupuna) on 21 January 1897 to the Secretary of the Education Department:

> . . . Perhaps, I need hardly draw the attention of your Department to the great need of educating the Maori children . . . and it is incumbent on the Government of this country, and all *true men and women*—in the interests of society . . . not to allow them to fall lower and lower and become a curse rather than a blessing to the community . . .

a quote from a Pakeha journalist:

> Existence of a disparity depends entirely on what is measured. Maori, they say, have values and competencies other than academic performance measures, which are not taken into account in disparity statistics. Maori

> might suffer a disparity in their School Certificate maths results, but Pakeha could be said to suffer a disparity in their lack of extended whanau (Gerritsen 1998:6).

and a story from the Maori author Patricia Grace, it goes something like this:

> A young Maori girl gave a morning talk about how she along with her kuia and koro spent some time killing "white butterflies" in their garden. The teacher did not like the story because, as she saw it, "white butterflies were beautiful".

True men and women from the Maori community have created Maori educational environments such as Kohanga Reo, Kura Kaupapa Maori and Whare Wananga. They have made a huge contribution to the country and now they want to know how well they are doing.

No reira ka mutu taku korero ki a koutou, tena koutou, tena koutou, kia ora koutou katoa

KUA MUTU

References

Gerritsen, J. (1998). Systematic failure. *NZ Education Review*, 6 October, 6.

Ministry of Education (1998a). *Assessment for success in primary schools*. Green Paper, May.

Ministry of Education (1998b). *Assessment for success in primary schools: Report of the submissions to the Green Paper*, November.

Simon, J.A. (1990). The place of schooling in Maori-Pakeha relations. Unpublished PhD thesis, Department of Anthropology, University of Auckland.

MINISTRY OF EDUCATION'S POLICY INTERESTS IN ASSESSMENT

Frances Kelly

As requested by the organisers of today's conference, the focus of this presentation will be the Ministry of Education's priorities in assessment.

I want to structure this by presenting an overview of the "desired outcome" state, and interweaving a description of how current policy directions are attempting to assist in reaching this outcome state.

In doing so I shall create a matrix overlay and apply this matrix to the policy areas of *Information for Better Learning* (MoE 1999) and *Achievement 2001* (1999). The key interest in this matrix overlay is in how the various elements operate in their connectivity.

MATRIX

	Collection	Collation	Analysis
Capability			
Credibility			
Clarity			

Capability: raising the capability of teachers to design or select and conduct appropriate assessment activities, and to interpret and act upon the information generated at the individual and aggregated level.

Credibility: quality assuring the assessment undertaken by teachers so that those receiving the information—whether students, parents, Boards of Trustees, employers, communities, government—can place a high level of trust in the information provided.

Clarity: focusing on the purposes of assessment and the balances among learning and teaching, qualifications and accountability.

Collection: the element of assessment which involves the systematic collection of information.

Collation: putting together information from more than one source.

Analysis: interrogating the information to determine future action.

With respect to the first square, then, capability and collection, the Ministry's interest, as evidenced in *Information for Better Learning* (MoE 1999), lies in having teachers who select or design and administer a range of assessment tools in order to collect useful information regarding the learning of their pupils. In order to do this, the teachers will need to be very familiar with the curriculum statements, and have access to a range of assessment tools, along with the ability to design tools of their own, with due regard for the technical validity of such tools, and an appreciation of the requirements for

administering the tools, and examining the resulting pupil responses in light of the rubric of expectations set out for the tools.

Current policy priorities to assist in this area include the commissioning of a wide range of new assessment tools to enhance the range of choices teachers can make, and the ongoing provision of professional development in the area of assessment, and to support the introduction of new curriculum documents.

The recently announced policy includes the commitment to a stocktake of current diagnostic tools in literacy and numeracy, the development of new tools to supplement these, the development of exemplars of student work across all seven learning areas, and the commitment to ongoing support of the ARBs and NEMP. In addition, the creation of new tests for literacy and numeracy will further add to the range of tools teachers will have at their disposal.

In the middle square of the top line is capability with respect to the collation of information. This can be seen in two dimensions. First, there is the collation of information from a range of assessments that together provide a rich picture of the individual learner, which informs the pupil, the teacher, the school and the parents. However, in collating information from a range of assessment tools, there needs to be consideration of the relationship of the various pieces, their weighting, the relativity of their validity, etc., so that the picture created is not skewed.

Collation of evidence at the individual student level is one element of collation. The other is collation at the level of the classroom or the school. Reports from assessment professional development contracts, funded by the Ministry, indicate that some schools are becoming increasingly adept at collating evidence in this way. The Ministry's interest lies in the expansion of such capability.

Finally, on the top row is capability with respect to analysis of the information gained. This, along with the development of new assessment tools, will be key focus areas for the Ministry's attention in the time ahead. For without capability in analysis, there is no assessment.

Take for example School Entry Assessment, or any other diagnostic tool. Using the medical analogy of diagnostic testing, unless the evidence is analysed, and the teacher is able to use the level of professional judgement required to put together the most appropriate programme to meet the students' needs, then conducting the test is not worthwhile.

Similarly at the class and school level, unless there is analysis of the collated data, and action is taken as a result of such analysis, teachers are likely to rely on the "affective" response to programmes, rather than the "effective" outcomes. Developing the habit of using assessment data to continuously review and refine teaching practices and programmes at the individual, the class, the school and the system level will give us the greatest opportunities to enhance student learning.

Across the second line is the combination of credibility with the collection, collation and analysis of data for learning. This element arises often in discussion around assessment. How credible is the judgement of a teacher? What moderation is there of that judgement? Given the importance of the outcomes for students, that is hardly surprising. Teachers and schools should be in a position to justify their choice of

assessment tools, and provide evidence to those involved that the judgements they have made are based on soundly collected data, from appropriate forms of assessment.

The parent pamphlet issued in connection with *Information for Better Learning* offered schools a chance to display that credibility. It posed the question: "How well is your child really doing at school?" In answer, it suggested that parents discuss this matter with the school concerned. It did not presume that government or one of its agencies or any particular piece of assessment would have the answer to that. The pamphlet did provide teachers and schools instead with the opportunity to demonstrate the credibility of their assessment practices. It is to be hoped that it will be picked up and used in that way.

In terms of policy directions, there are a number of elements of the recently released policy which will contribute in raising credibility. First, the development of a range of new assessment tools will assist by allowing teachers to access the level of professional credibility of the assessment experts in this country, as opposed to being too often in the situation of having to "invent the wheel" in each situation. Increasing use of the assessment resource banks, and the widened range of diagnostic tools, will also assist here.

Secondly, the development of exemplars of student work will assist in providing benchmarks for judgements across the learning areas, and across the curriculum levels.

Thirdly, the structure of the pilot of externally referenced tests will enable us to look across the school system at the data available for analysis, and at how such analysis can best be supported.

The final square relates to clarity. Wynne Harlen, in her keynote address to the NZARE conference in 1998, spoke of formative and summative assessment as being separated not so much by the type of assessment as by the approach to the analysis undertaken. Having analysed data for formative purposes, it is not appropriate to treat that analysis summatively, but the original data should remain available for a second round of analysis for other purposes.

What does this mean in terms of collection, collation and analysis? There is a need for awareness of the multiple uses of the data collected in assessment to ensure that assessment *per se* does not dominate the classroom or school practice; and there should also be clarity about the purposes for which data is being collected, clarity about what is being collated, and clarity about the process and point of analysis.

A professional stand needs to be taken regarding the collection of accounts of pupil performance for which no further use is envisaged following their recording. This should not be confused with assessment for accountability. Schools are accountable for effecting the greatest possible progress in the learning of each individual and cohort of pupils. Assessment contributes to this only when the accounts of pupil performance change pupil, teacher, school, or system behaviour. Simply recording and reporting pupil achievement is not sufficient. Nor is finding some way of calculating the value added factor for which the school can claim credit. Unless the school can show that following the recording, reporting and calculating, changes were made which enabled student learning to be deeper and more assured, then the New Zealand school system will have underused the power of assessment to effect better learning.

In this context, there is a need to ensure shared understanding of the government's recent announcements in *Information for Better Learning*. In particular, correction is

needed to NZCER's introduction to the conference, as shown in the programme. Government has not announced the introduction of whole cohort testing programmes, as the programme claims. It has announced its intention to undertake a pilot study of externally referenced tests.

This pilot has three parts. It will involve a 10 percent sample of schools. Part one sees all schools in the sample receiving professional development assistance, focused on the use of assessment information in decision making and communication, at the individual teacher, class and school level.

Part two sees half of the schools in the pilot including newly developed externally referenced tests of literacy and numeracy for their pupils at years 5 and 7, along with their current range of assessment methods. The other half will continue to use only their current range.

Part three is an evaluation of the use made of the assessment data, focusing on the value added of the externally referenced tests, and the conditions under which such value can be maximised.

I wanted to spell this aspect of the policy out in some detail, as it is people attending a conference such as this, those interested in assessment and those interested in research, who may be seeking involvement in the design and development of tests, the provision of professional development and the evaluation study.

In concluding this part of the presentation, I want to look at a couple of the statements about assessment which I hear tripping off tongues with more frequency and less thought than is probably useful.

During and since the consultation on *Assessment for Success in Primary Schools*, I have heard too often that pen and paper tests have no value. This conference is sponsored by NZCER. This body develops a range of pen and paper tests using multiple choice responses which are used throughout New Zealand schools, and were frequently identified as being of significance in the consultation responses. In the assessments used in the National Education Monitoring Project, a considerable number involve pen and paper—indeed that is how they are described in the reports. The assertion that pen and paper tests have no value has become a mantra for some, and needs more careful consideration.

The second is that assessment doesn't improve student learning. Often this has been stated as "no matter how often you weigh a pig, you won't fatten it". In terms of this presentation, that underestimates that which has been defined as assessment. The statement has its focus on merely the collection and perhaps some element of collation of data. It certainly does not include analysis. Yet it is the power of that analysis of data which researchers look to constantly in their quests for understanding. Collection and collation of assessment information does not of itself improve student learning. Action taken as a result of the analysis gained from assessment does.

I want now to turn to the work being undertaken in assessment for qualifications. Given the work currently being undertaken by the Qualifications Development Group and the widespread consultation currently operating throughout the country regarding the National Certificate in Educational Achievement (NCEA), it is appropriate to look at that policy development using the same matrix.

It is now almost 12 months since the government announced the new qualifications. The impetus for the introduction of the NCEA came from a period of some confusion, with clarity, credibility and capability with respect to assessment at its heart. The policy behind the NCEAs now in development is to make the best use possible of the gains in these areas over the past ten years, while recognising the needs for surety from the users of the qualifications, whether these be students, schools, higher education institutions or employers.

I shall begin this time with the bottom row, considering clarity in collection, collation and analysis.

The policy analysis behind this work began with a clarity of focus on the various purposes for which the qualifications would be used, so that appropriate decisions could be made about the assessment practices and procedures which would inform them. The numbers of school leavers using each of the three levels of qualification, their destinations, and the needs of the organisations and/or individuals to whom they would present their qualifications were considered.

With respect to collation, consideration was given to the need for clarity in the systems established for reporting the outcomes of the qualifications, so that there is a balance between the detailed account of what the student can do, used frequently for finer judgements, and the documentation which tells the student and other stakeholders that a meaningful qualification has been gained. That is why the format for reporting is being included in the current round of consultation, to ensure that the models proposed in the initial policy document can meet the needs of the various stakeholders.

Thirdly, with respect to collection, the policy analysis focussed on clarity in the selection of appropriate assessment activities, making them sufficient to allow pupils the opportunities to demonstrate that which they know and can do, but not so suffocating that they overwhelm the teaching programme. In addition, there needs to be clarity in the process for assurance that the assessment activities are valid for the achievement objectives, and enable distinctions between credit, merit and excellence to be made.

To this purpose the development of the qualifications includes the development of assessment exemplars, which will be available for teachers and schools to use in assessing against the achievement standards. This development has already started, as part of the task of the various achievement standard development groups.

The middle row focuses on credibility. In considering policy areas relating to credibility in assessment, there is a tension to be managed between the credibility of the assessment tool, and the credibility of the independence of the assessor's judgement. This is evidenced for example in the long standing tension between examination systems that use a small set of assessors removed from the teaching programme to maximise the independence of their judgement, and the wider range of appropriate assessment activities that best assess particular knowledge and skills.

The NCEA policy seeks to maximise the contributions of both. Having developed the achievement standards in a learning area, the panels (and the consultation process) seek to identify those which, with a minimum of 50 percent, can appropriately be assessed through external assessment (including examinations), and those which can most appropriately be assessed through internal assessment, using a wider variety of tools.

This will not therefore involve using the external assessment to moderate the internal assessment. Nor will it, as the present Sixth Form Certificate procedure does, moderate the outcomes of one year's study on the basis of the performance of the same cohort 12 months earlier in external examinations. Given that in the development phase, the focus has been on the credibility of the assessment process to assess the specific achievement standard, neither would be appropriate.

There will, though, be a rigorous moderation system for the internally assessed components. Just as the writers of examination papers manage the constraints of that form of assessment, so moderation operates as a quality assurance procedure to ensure that for the users of the qualifications, the credibility of the internal assessment records matches that of the external.

The proposal under consultation at present looks to an accreditation/audit model, with strong institutional sanctions where the choice and/or conduct of assessment does not meet the high standards expected. The accreditation/audit cycle will follow a regular cycle, with the provision for spot audits, where an analysis of the overall outcomes of internal assessment does not fit within toleration limits. There may well in some cases be excellent reasons for this. The particular strengths of individual teachers or groups of students within schools will always mean that such statistical analysis of aggregated data can only be used as an indicator of interest.

Finally, on the top row sits capability. Capability with respect to assessment in the senior secondary school for qualifications purposes has developed considerably over the past 10 years. Underlying the use of unit standards in 330 out of 360 secondary schools lies a serious long term enhancement of professional capability in the collection and collation of assessment information, which should not be undervalued. It is a very long time (if ever it was the case, as anything more than an urban legend) since UE accreditation was a sure thing for the favoured few sporting greats. Whether through involvement in assessment in Art, Music, Technology, or the whole range of subjects in Sixth Form Certificate, the sector capability in assessment should be recognised as a considerable asset.

This needs to be remembered when critics outside the system attempt to create doubt. New Zealand teachers working in the senior secondary schools have long had experience of moderation of their assessment, requiring them to be able to verify the authenticity of the work they assess, the process by which they undertake that assessment, and the shared understanding of the standards they assess against.

This is not, however, to say that the credibility of qualifications can ever be put at risk by merely assuming this capability. The need for moderation of assessment is an ongoing one.

These then are the component blocks of the Ministry's interests in assessment: capability, credibility and clarity, in collection, collation and analysis.

Finally, I want to turn to three areas of potential for future research work in this area.

First, there is a need for ongoing research which informs the design and development of high quality assessment tools; research into current practices of teachers and schools; and research into professional development practices which will most usefully enhance current practices.

Secondly, there is a special need for research into the developmental processes of becoming a second language learner and bilingual adult, so that appropriate diagnostic tests and other assessment tools for teachers and students can be developed. Without a clear understanding of the destinations and possible journeys of the learner, the data collected in assessment cannot be analysed. In particular, in New Zealand, this work is critically linked to the supporting the work of those in Maori medium and bilingual education. The development this year of a Maori language proficiency test for year 5 students, and the policy provision for a similar test for year 8 students, along with the publication and professional development to support the use of these tests, will provide a base for further work. So too will the inclusion of assessment of students in Maori medium education in the Year 8 NEMP study. But there is need for further basic research in this area.

A further challenge comes in the integration of assessment and information and communication technologies. With the growing power of connectivity, there needs to be considerable work undertaken in looking at its potential, not only because it widens the possibility for assessment design, but also because it is fundamentally changing the ways pupils learn. In Victoria, for example, the provision of internet assessment allows students changing schools or coming into the state to access interactive assessments online. The interactivity of this system quickly responds to the ability level of the student. The CD-rom version of the school entry assessment developed at Durham University offers similar interactivity for 5-year-olds. This means that the nature of the questions the student faces is determined by their ability to answer a previous set. The intent is to minimise the negative impacts for a child who struggles from the start, while ensuring that a bright child does not lose interest in the challenge.

There is no use in pretending that capability, credibility and clarity in collection, collation or analysis, as discussed in this presentation, is universally present throughout the teaching profession and schools in New Zealand. The challenge for the Ministry is to develop policies which will support its development, for those working in education research to explore the realities of the present situations, and the potential for where it could be, and for teachers and schools throughout the country to maximise the potential of what they do now, and look constantly for opportunities to improve.

Above all we must avoid painting ourselves into indefensible positions.

ACHIEVEMENT 2001

Michael Irwin*

Policy Analyst
New Zealand Business Roundtable

* I am grateful to Glennys Allen, Cedric Croft, Cedric Hall, Roger Kerr and Tony Locke for valuable observations and suggestions on an earlier version of this paper.

Introduction

Achievement 2001, the policy for school qualifications, was announced by the former minister of education, the Hon Wyatt Creech, on 5 November 1998. There had been no prior public consultation on school qualifications except in the context of the submission process on the government's Green Paper on the National Qualifications Framework (NQF), which was released in June 1997.[1]

There was only a page and a half specifically on school qualifications in the Green Paper. It declined to be drawn into a debate about the merits or otherwise of school examinations and said that criteria for the registration of qualifications should not favour one assessment practice over another. The announced policy would establish an "inclusive" and "comprehensive" framework which would be "settled and stable". However, registration criteria for qualifications would include the use of "explicit learning outcomes" and "credit values" towards the National Certificate of Educational Achievement (NCEA).

The policy decisions as announced in November 1998 followed the general lines of the Green Paper. According to the announcement:

- *every* student will study towards the NCEA, and *all* learning in the senior secondary school will generate credits towards the NCEA;
- there will be four levels to the NCEA (typically fifth, sixth and seventh forms and scholarship);
- in conventional curriculum subjects credits can be gained by a mix of internal and external assessment which would be against "Achievement Standards" that will be developed by expert groups and will set the standards to be met. Achievement Standards will also recognise merit and excellence and at least half are to be assessed externally;
- School Certificate (SC) and Bursary would remain and form the external assessment component of the NCEA;
- Unit Standards will be employed for non-conventional subjects at levels 1 to 3 with the notion of "merit" not being ruled out; and
- credits would be attached to Achievement Standards and some Unit Standards towards the award of the NCEA.

In announcing this policy[2] the minister described much of previous debate about the NQF as "very arid and rather pointless" and urged that it was time to "bury" it. He saw his new policy as a "well constructed middle path" and, while some "extremists" would still argue about it, one which the "great bulk of the sector [would] back".

At the risk of being seen as "extreme", I have problems with important aspects of the announced policy and propose to raise a number of difficulties. I would also point out that the majority in any cause can be wrong—if it is true that the "great bulk" of the school sector support Achievement 2001 then that fact alone doesn't make the case for it.

---

[1] Creech, W. (1997), *A future qualifications policy for New Zealand: A plan for the National Qualifications Framework*, Wellington: Ministry of Education, June.

[2] Creech, W. (1998), Launch of Achievement 2001 – School qualifications for 16 to 19 year-olds, speech at Rangiora High School, Rangiora, 5 November.

But I have no idea what measures if any have been taken by the ministry to ascertain the depth of support.[3] I would doubt if, in any case, many in the sector have been sufficiently well informed about it to enable them to make a responsible decision one way or the other. There is no official paper that I know that has been made publicly available and draws on local and international research, analyses the problems, identifies and evaluates the various ways of addressing them, and explains why certain options were chosen and not others and how the inevitable trade-offs were resolved.

I am all for a robust and enduring schools qualifications system but consider that it is far too important a matter to be decided by some sort of majority vote. Of course teachers must be actively involved and their knowledge and experience drawn upon. They must be persuaded of the merits of whatever the final decisions are to be if a policy which depends on their cooperation is to work. But teachers and others will want to be satisfied that whatever is proposed has been developed from first principles, from a sound theoretical understanding of assessment and of the pragmatics of the classroom including the many pressures on teachers and schools, and as a result of acknowledgment and credible resolution of the tensions and trade-offs that are inevitably involved in the construction of an assessment system.

I have a strong sense of déjà vu. Not long ago a totally unit standards-based NQF was being promoted with evangelical fervour as the indubitably correct solution to our educational problems. I also recall being told by a ministerial speech writer that I was the object of a reference in a speech by a previous minister of education to a "Custer's last stand" being made against his NQF policy. I had never seen myself as a swash-buckling military figure before! In April 1996 it was announced that the NQF would be "broadened" to include whole qualifications. This and other changes effectively meant that the central unifying idea of unit standards had been abandoned, though they were misleadingly portrayed as part of the "natural evolution" of the NQF.[4] And here we are again being told that a system based entirely on achievement standards and unit standards is indubitably the best school qualifications system and only "extremists" might dare to differ.

The initial ministerial announcement and accompanying documents contained a number of ambiguities. But I think at the time most of us were too punch drunk by all the policy developments and debates, including assessment debates, to pay close attention. Perhaps we saw in the announcement whatever we wanted to see—norm-referenced exams for those who wanted them retained and standards-based assessment for those who preferred that form of assessment. "Feel-good" statements that Achievement 2001 was a "middle path", "comprehensive", and employed the "best of what we have now" might have helped to anaesthetize our critical faculties. In any case it was close to Christmas, and there were other things to think about!

---

[3] A nationwide survey of secondary school teachers by a group calling themselves Concerned Teachers, reported in *The Education Weekly* of 11 October 1999, indicated that 87 percent of secondary teachers who responded to a national survey wished to retain School Certificate. A spokesperson for the group said that the survey was carried out "to verify earlier feedback showing major concerns about 'Achievement 2001' ".

[4] For discussion see Irwin, M.D.R. (1997). The National Qualifications Framework: Where to now?, *ACCESS*, 16(2), 14–26.

Two Initial Observations

Lack of Consultation

My first observation is that the scheme as a whole has not been subject to public consultation. I understand that there are various advisory groups and, of course, expert groups for particular subject areas, but they are involved in implementation issues: the basic design had already been set.[5] Certainly in some, but not all, respects the scheme is similar to that outlined in the Green Paper, but the discussion in that Paper was extremely brief and I am not aware that problems raised about it have been answered.[6] I think the lack of consultation on the scheme is bringing its own difficulties, though I have to say that my experience with the consultation process used in the curricular area does not encourage me to think that consultation over Achievement 2001 would have led to any significant change.[7] Presumably the calculation was made that the scheme as announced would appeal to all reasonable people and further consultation would simply be a waste of time. After all, only "extremists" would oppose it. As it is, serious concerns remain and important details have to be worked out, and the devil is, as has so often been said in other contexts, to be found in the detail.

The Speed of Development and Implementation

Development is now happening at great speed and schools must be preparing to introduce the new level 1 Achievement Standards. So my second observation is the incredible speed proposed for the development of the Achievement Standards and their introduction. No trialling is proposed; in effect, the whole 1999 Third Form cohort is to be an experimental group for each of the three years of the scheme's introduction, 2001–2003. It suggests quite astonishing confidence in the new system—a confidence which is certainly not justified by the analysis in any of the papers I have seen. The proposed national testing for primary children is to be trialled, so why not Achievement 2001, which is surely at least as significant?

Quite apart from any design defects, the sheer extent of the development required is such that some inconsistency in policy application, unevenness in the interpretation of what is required and frequent revisions would seem to be inevitable given the compressed timeframe and lack of trialling. I have on my desk a pile about four inches high of Achievement Standards, but they are only for 15 subjects and only for level 1. With standards required at three grades for each of the four levels for each Achievement Standard and from 5 to 8 Achievement Standards per subject, it is easy to see how large

---

[5] A questionnaire, to be returned by 28 October 1999 and sent to all schools by the New Zealand Post Primary Teachers' Association (PPTA) and the Ministry of Education, seeks comments on aspects of the proposed scheme—not on the scheme as a whole.

[6] *See* Smithers, A. (1997), *The New Zealand Qualifications Framework*, Auckland: The Education Forum, November; Irwin (1997).

[7] For discussion of the consultation process on the draft curricula, see Irwin, M.D.R. (1999a). The curricular reforms – are they taking us forward or backwards? Paper presented to a course on contemporary education policy, Victoria University of Wellington, 17 July.

the task is.[8] If there are about 30 "conventional" subjects we could end up with some 2500 sets of standards. Perhaps a third of these will be required at the beginning of the 2000 school year, "so that teachers would have a year to consider programmes that would prepare students for assessment against the standards in 2001."[9]

Two Underlying Problems

The Concept of a Framework

Achievement 2001 represents yet another attempt to put flesh on the concept of a qualifications "framework". In 1991, in the very early days of the construction of the aborted NQF, John Codd, Don McAlpine and Jenny Poskitt wrote as follows:[10]:

> The concept of a "framework" for national qualifications . . . is one which has obvious appeal to policy-makers. It represents a unified, logically planned and systematic approach. . . . In theory, it is a concept with considerable plausibility; in practice, however, it may be extremely difficult, if not impossible, to achieve.

They went on to point out a number of tensions and conflicts, including those between:

- the standardised outcomes required for a framework and institutional sources of energy and diversity which standardisation might restrict or suppress;
- the inevitable constraints of any framework and the desirability of flexibility and diversity;
- the adoption of a single assessment methodology which is desirable for framework purposes and the fact that one assessment methodology cannot meet all certification requirements; and
- the requirement for a common structure of levels and credits which form a necessary basis for a framework and the very different types of knowledge and skills to be fitted within it.

To my mind this was very perceptive and much of what they said back in 1991 is as true today for Achievement 2001 as it was for the then emerging NQF. The paper listed some of the tensions and conflicts which need to be addressed openly and squarely in designing a framework, whether one for all qualifications, as with the NQF, or only for a particular group of qualifications, as with Achievement 2001. They reflect epistemological constraints, the limits of assessment technology and the diverse purposes which qualifications have to meet. They have been commented on in many papers on the NQF and in submissions on the Green Paper. The basic argument for uniformity is simplicity. But if we seek to apply a uniform approach to assessment over a wide range

---

[8] There is some debate whether "Scholarship" will be a separate level or will be awarded for superior performance at level 3.

[9] National Certificate of Educational Achievement – background for consideration in consultation, undated Ministry of Education paper, p. 3.

[10] Codd, J., McAlpine, D., & Poskitt, J. (1991), *Designing the framework – a position paper on problems and issues in the restructuring of national qualifications*, Palmerston North: Education Department, Massey University.

of learning we end up with all sorts of distortions. Certainly what we were offered in the NQF and now find in Achievement 2001 is far from simple.

In the final analysis a judgment has to be made about where the balance should be struck between the perceived benefits of a framework in terms of uniformity and the costs of trying to force diverse material and assessment purposes into one structure which cannot accommodate everything without the costs of at least some distortion. And some of those costs arise from the implicit denial that there is a valid distinction between academic and vocational learning—that learning can be an end in itself.[11] I find very little hint of the existence of this fundamental problem in the official papers available to me.

We seem to be going through a stage of our educational history characterised by a search for frameworks, clear learning outcomes, transparent standards and the like. I have to say I am bothered by it: we are collectively suffering from a severe case of *hubris*—the mentality that places us over nature and seeks to reduce all knowledge, understandings and skills to numbers, to levels, to credit values and so on, and which appears to deny the mystery, the uncertainty and the qualitative aspects of our human existence and of our search for the truth about ourselves and the rest of the physical world. We are moving, dangerously in my view, from a view of education as open with unpredictable results to a reductionist view of education as closed and with predeterminable results.[12] If education is seen as closed and predictable, then we place it at the mercy of the social engineers; we have moved from education to indoctrination.

Of course, the pragmatics of the educational enterprise require some, often arbitrary, ordering of material. But we should always be aware of epistemological constraints and the limitations of assessment technology. This involves facing up to the tensions and trade-offs involved, being explicit about our judgments and being prepared to defend them. I anticipate that sometime in the future we will find that education is really much more messy than we thought and resistant to the nice neat categories and distinctions which our search for frameworks assumes exist. The trouble is that we can, in the meantime, do a lot of damage as we lop and chop to make education fit preconceived procrustean beds. In my view, our framework fixation has led to some poor judgments and is still doing so.

The Search for Both Uniformity and Inclusiveness

There is one particular dilemma which arose in the case of the NQF and which arises just as starkly with Achievement 2001. The dilemma,[13] in brief, lies in the tension between:

- a framework emphasising uniformity and thus interchangeability of component parts; and
- a framework emphasising comprehensiveness and inclusiveness.

Both types of framework have their problems and limitations. The original NQF was of the former type: the rigorous specification of unit standards as the common building

[11] A point made by John Codd (1996). Higher education and the qualifications framework: A question of standards, *Delta*, 48(1), 63.

[12] Irwin, M.D.R. (1999b). From virtues and vices to passionate values. Paper presented to the Independent Christian Schools' Fellowship, Auckland, 19 August, 9–11.

[13] See Irwin (1997).

block was aimed at maximising comparability, portability, "seamlessness", and the like. The "broadening" policy announced in April 1996 and elaborated in the 1997 Green Paper proposes the latter type, though misrepresented as a natural evolution of the former and one that maintains all its assumed advantages. The problem is that there are trade-offs to be taken into account. The more uniform the system, in terms of its structure and methodology, the less inclusive it can be in terms of its ability to embrace different methodologies.

Achievement 2001 also wants to have it both ways. The Green Paper, as we have seen, proposed a "comprehensive", "inclusive" framework and one which can accommodate a variety of assessment methodologies, and yet it immediately went on to talk about "explicit learning outcomes" which clearly points to a particular type of assessment. The minister, in his speech announcing Achievement 2001, claimed that the new policy was "a well constructed middle path" which, like Tony Blair's impenetrable "Third Way" in politics, steers a successful course between the divides set up by what are portrayed as the "arid" arguments of the past. But the chosen "middle path" aims at uniformity and will necessarily exclude other "paths", thus limiting comprehensiveness and inclusiveness.

The speech also claimed that the new "system brings together the best of what we have now". We are not told what is the "best" of what we have now, the basis on which the evaluation and selection were made, and how these superior components are to be assembled. But whatever combination is in mind will necessarily exclude other combinations. Moreover, the "best" regime for one purpose may not be the best for another: the best assessment methodology for selecting entrants to a medical school is not necessarily the best one for certifying competence in typing. In short, while it is claimed that Achievement 2001 has resolved the tension between uniformity on the one hand and inclusiveness on the other, the words actually used in making this claim make it clear that the dilemma remains unresolved.[14]

The Assessment Methodology

But perhaps the starkest and most serious problem is that of assessment methodology. As we have seen, the 1997 Green Paper was ambiguous. It emphasised inclusiveness and said that NQF registration criteria should not "be used to decide in favour of one assessment practice over another". It went on to say that "where options are available, secondary students could choose how they wished to be assessed". This seemed to mean that various forms of standards-based assessment (including standards- and achievement-based) and norm-referenced assessment, including forms of scaling, were all open for consideration and selection according to the material to be assessed and the purposes of assessment. But then in stating that registration criteria should include "explicit learning outcomes" and "credit values" it seemed to contradict the emphasis on "inclusiveness" and point to some form of standards-based assessment within a unified system.

---

[14] A consultative forum established by the Secretary of Education met on 7 and 8 September to consider issues that had emerged. It recommended, *inter alia*, that the NCEA be "As **Inclusive** and **Flexible** as possible . . . "and that it be **Coherent** (a unified system)" (emphases in original). The forum did not appear to recognise the trade-off between inclusiveness and uniformity.

The minister's speech[15] launching Achievement 2001 in November 1998 went some way towards resolving the ambiguity: it seems clear that "inclusiveness" had been rejected in favour of uniformity. There would be *one* "middle path" for *all* students and *one* "unified" form of assessment. *Every* student was to study towards the NCEA and *all* learning in the senior secondary school would generate credits towards it. Moreover only "skills and abilities" were to be assessed—there was no mention of knowledge, understandings and such like, but this was before we were introduced to the heady notion of the "knowledge" economy. A media statement issued with the minister's announcement advised that the minister said "that teachers would no longer have a dual system of assessment for Unit Standards and exams". One of the government's objectives in "unifying" assessment was to "remove the status distinction between the standards-based and traditional examination approaches to qualifications assessment."[16]

But there were still some ambiguities and unacknowledged tensions. For example, the minister stated that for "regular" subjects SC and Bursary "will remain and form part of the external assessment component of the National Certificate". But how could examinations, which are norm-referenced and employ forms of scaling,[17] remain along with Achievement Standards and Unit Standards within a "unified" qualifications system employing a single form of assessment? I note in this regard that Achievement Standards are to be "more consolidated, less atomised than Unit Standards" which is fine as far as it goes, but the different degrees of specificity may require different forms of assessment.

An official brochure issued at the same time as the announcement[18] said that "Examination marks will not be scaled" and as Achievement Standards are to replace examination prescriptions, it is hard to see how SC and Bursary would in fact "remain" in any recognisable form. Moreover, success in some Unit Standards is to contribute credits to the NCEA: how is the assessment methodology for them to be unified with that of Achievement Standards and examinations? Further, the brochure assures us that Achievement 2001 would provide "results that show how well students have done against each other and against national standards" which seems to imply norm-referenced as well as standards-based assessment—unless, that is, relative performance is simply crude differentiation by three grades (credit, merit and excellence). Again, how does this square with the single "unified" system that "brings together the best of what we have" and avoids a "dual system of assessment"?

A ministry memorandum[19] advises that "Achievement 2001 . . . [is] to develop a "unified qualification" that [*inter alia*]:

---

[15] Creech (1998).

[16] Qualifications Development Group (QDG) (1999), *Update*, 15 June.

[17] There are various forms of scaling including inter-marker scaling to even out differences in marker toughness and inter-subject and inter-year scaling to even out differences in subject difficulty and in the difficulty of papers between years. Inter-year scaling is said not to be undertaken for the SC. However, the NZQA prepares a provisional marking guide, marks a random set of scripts, checks the number of students who would achieve each grade, and then adjusts the marking guide as necessary to avoid major variations between years in the percentages in each grade. Thus a similar distribution of marks is maintained over time. Given the impossibility of setting questions of equal difficulty each year this practice seems entirely sensible, indeed essential.

[18] Ministry of Education and New Zealand Qualifications Authority (1998), Achievement 2001 – Qualifications for 16 to 19 year olds, Ministry of Education, Wellington. (Undated but presumably issued in November.)

[19] Ministry of Education (1999). An internal memorandum to the Secretary of Education entitled The National Certificate of Educational Achievement – paper 1. (Undated).

- may be gained by internal assessment (performance) and external assessment (including written examinations and externally assessed portfolios and projects); and
- unifies and removes the status distinction between the standards-based and traditional examination approaches to qualifications assessment.

It is also claimed to be "simple to understand", "easily manageable by teachers", and that it "reports information that is meaningful and useable" and which is "credible with the stakeholders". We are not told quite how all these very desirable objectives are to be achieved within a single, unified system. They are, however, to be achieved within a standards-based assessment system. It is hard to avoid the conclusion that the so-called "unification" of standards-based and traditional examination approaches is to be undertaken by expanding the former and abolishing the latter. For example, the authors of the paper:

- note that the government decided that "excellence and merit [would be] recognised through grading of achievement against standards";
- consider that "Qualifications are valuable only if they are meaningful and credible—that is, that the meaning of certification is transparent, and that the holder *really* knows or can do what is attested" (emphasis in original); and
- predict that the White Paper on the NQF will "widen and strengthen the NQF" and "continue" its "principles" and require "high quality" qualifications to "give accurate information about what a student knows and can do".

All this is very much unit standards "talk"; it's hard to detect any elements of a "traditional examination approach" that remain in the "unified" system. Perhaps the view was that the credibility of standards-based approaches could only be assured by removing the competition of norm-referenced examinations.

It seems clear that a form of standards-based assessment—achievement-based—is to be adopted. However, we are not told precisely what it means and how it differs from competency-based assessment which is the form of standards-based assessment employed for unit standards. Is it simply that the achievement–based assessment is graded whereas competency-based assessment is not? Will there be any conceptual difference as regards assessment methodology between Achievement Standards and Unit Standards if the latter adopt a merit level as has been raised as a possibility?

Terry Locke has analysed the transformation that takes place when the mode of standards-based assessment changes from an achievement-based one to a competence-based one as in Unit Standards.[20] Essentially the transformation involves the replacement of qualitative judgments by technicist measurements. Such a change clearly poses considerable problems for assessment in a subject such as English, as, indeed, in other

[20] Terry Locke has set up a website in connection with his English Study Design project to make available information on Achievement 2001 especially as it relates to English (http://www.tmc.waikato.ac.nz/support/artslang/esd/2001forum.html). The URL for his article Achievement standards: What's in a name? is <http://www.tmc.waikato.ac.nz/support/artslang/esd/achievestand.html>. The URL for his article Lacks of Achievement 2001 is <http://www.tmc.waikato.ac.nz/support/artslang/esd/2001/EIAArticle.html>. See also Locke, Standards-based assessment in English: Take 3, *Waikato Journal of Education*, forthcoming.

"conventional" subjects at the senior secondary level. Moreover, Locke shows that the experts panel for English had (as at August 1999) moved towards the development of Achievement Standards in the Unit Standards competence-based format with quantitative "clip-ons" for the merit and excellence grades. He argues that in the draft English Achievement Standards for level 1, the proposed "basis for awarding merit and excellence . . . is too inadequate to be seriously countenanced" and has serious pedagogical implications.[21] I am not aware of any ministry response to his very detailed analysis, which would appear to have considerable significance for its work on policy for school qualifications.

It would be useful to know why achievement-based assessment was rejected by the NZQA in 1993[22] and what subsequent research has led to its adoption now.[23] If, as Locke's work suggests, the development of at least some of the Achievement Standards in fact draws on a competence-based model rather than an achievement-based one (as, of course, might have been inferred from the name of the project, *Achievement* 2001), how does this square with the ministry's acknowledgement of difficulties with Unit Standards which also adopt the competency-based approach? However, the ministry has provided no analysis of these difficulties which, had it been undertaken in a rigorous professional way, might have led to very different policy prescriptions to those embodied in Achievement 2001: the ministry's policy papers[24] rely more on self-congratulatory assertion and vacuity than on analysis and research.

The ministry's cavalier attitude to analysis and research and its appointment of many Unit Standards proponents to its expert groups suggest a continuing commitment to a Unit Standards-based qualifications framework. Its chief executive, Howard Fancy, has advised other departmental chief executives that the concerns about the National Qualifications Framework "centre less around the intent and general concept of the framework but more around implementation issues and its long term management"—this is simply asserted and is not the conclusion to rigorous analysis.[25]

On the issue of research, Warwick Elley observed in 1993 that:

> . . . it should be stressed that educators in many other countries have tried to develop clear stand-alone standards in general subjects at the upper secondary level, but none have succeeded.[26]

---

[21] Locke, forthcoming (see note 21).

[22] NZQA (1993), *Framework Update*, August, 5.

[23] See Peddie, R. (1995). Competency, standards, merit and excellence. In R. Peddie & B. Tuck, *Setting the standards*. Palmerston North: Dunmore Press.

[24] Papers obtained under the Official Information Act include the ministry's reports to its minister dated 23/3/98 (entitled Qualifications for 16 to 19 Year Olds), and those of 17/7/98 and 20/8/98 (both entitled Qualifications for Young People Aged 16 to 19 Years) and the minister's undated paper to the Cabinet Committee on Social Policy entitled Qualifications for Young People Aged 16 to 19 Years.

[25] Fancy, H. (1997). Memorandum to OCETE Chief Executives entitled "Senior Secondary School", 9 December.

[26] Elley, W.B. (1993). The NZQA agenda unfolds. *The Christchurch Press*, date unknown, cited by Croft, C. (1993), in *The conflicting world of standards-based assessment*. Paper delivered at the 15th Conference of the New Zealand Association for Research in Education, Hamilton, 2–5 December.

Cedric Croft commented in 1993 and 1994[27] to the effect that there was then little expertise in achievement-based assessment available within New Zealand and that little research into it had been undertaken. Some of the research that had been undertaken was not encouraging.[28] The only recent trial and professional evaluation of an achievement-based system at the senior secondary level has been that conducted in the context of the English Study Design (ESD) project, coordinated by Terry Locke and funded by Waikato University's School of Education. The ESD project was set up to provide a coherent, rigorous, senior English programme, using standards-based assessment and national moderation procedures, as an alternative to Unit Standards.[29] It is not clear what, if any, lessons the ministry has drawn from it.

In 1995 Warwick Elley, Cedric Hall and I wrote an analysis of Unit Standards which concluded that a basic fallacy was that "one type of assessment (a particular form of criterion-referenced assessment) is suitable for assessing the myriad combinations and levels of skills and knowledge to be found in education and training courses and programmes".[30] Much that was said about Unit Standards would seem to apply to Achievement Standards. In 1996 Warwick Elley wrote an article discussing the particular problems that arise when trying to fit academic courses into the NQF. In it he drew attention to a number of difficulties regarding the definition, assessment and moderation of standards in academic school subjects which, he pointed out, are "multidimensional and irreducible to a 'ladder of mastery' ".[31]

Where are the ministry's responses to the problems raised in these and many other papers?[32] In what respects were these and other authors mistaken in their analyses? Is there some research of which they were unaware and which is such that the ministry can now confidently recommend the introduction of a standards-based assessment approach for most, perhaps all, school-based qualifications?

Particular questions arise about how exactly the external assessments are to be conducted. There is to be no scaling; assessment is to be against standards not against the performance of other candidates;[33] Achievement Standards will replace prescriptions; and results are to attest to what the candidate "really" knows and can do—the intention seems

---

[27] Croft, C. (1993); (1994), *Standards-based assessment – what do we know? What should we do?* Paper delivered at the 16th National Conference of the New Zealand Association for Research in Education, Christchurch, 1–4 December.

[28] For example, Gilmore, A. (1991), *An evaluation of the sixth form certificate moderation trial*. Christchurch: Education Department, University of Canterbury.

[29] Hall, C. (1999), *Evaluation of the year 12 English study design trial 1998*, Victoria University of Wellington, Wellington, February; Locke, T., & Hall, C. (1999), The 1998 year 12 English study design trial: A standards-based alternative to unit standards, *New Zealand Annual Review of Education 8*, 167–187, Wellington: School of Education, Victoria University of Wellington.

[30] Irwin, M., Elley, W., and Hall, C. (1995), *Unit standards in the National Qualifications Framework*, Auckland: Education Forum, May.

[31] Elley, W. (1996), Unresolved Issues in Fitting Academic Courses into the Qualifications Framework, *Delta*, 48(1), 67–76.

[32] A recent issue of *ACCESS* was entirely devoted to articles about the National Qualifications Framework. *ACCESS* 16(2), 1997.

[33] However, standards usually have a normative basis ie they are based on what it can reasonably be expected that a student knows or can do at the particular stage or age in question. In the joint Ministry of Education/PPTA questionnaire recently sent to all secondary schools it is proposed (at Issue 3: Level of Difficulty) that standards for Level 1 credit be set so that a majority (70–80 percent) can be expected to achieve them. This clearly accepts the normative basis for standards. However, setting the "norm" at a very low level of achievement to ensure a high pass rate will be at the expense of making a "credit" virtually meaningless and failure totally degrading.

clear enough. However, a complication has been introduced by a recent announcement that "the new qualifications system will give students a percentage mark".[34] The consultative forum recommended that "Externally assessed standards are to be reported with raw marks and a percentile ranking for the mark". It is hard to understand how such a requirement can be met within an SBA system unless a great many more standards are to be produced to differentiate much more finely between students than the three grades (or four grades if failure is counted as a "grade") system allows. Perhaps we have here some political maneuvering in the face of opposition to the removal of external examinations—a move away from uniformity back to inclusiveness. But if so, what does this do to the framework?

It is also hard to see how the system could provide inter-year comparability, which is essential if the NCEA is to be credible and useful, without inter-year scaling, and there is to be no scaling of any sort. Moreover, if there are, in essence, to be two assessment systems, one for internal and another for external assessment, how does that fit with the decision to do away with dual assessment which is said to be one of the problems with our present system and one which Achievement 2001 will remove? We could ask the same question about the statement that the "new National Certificate [is to] be buttressed by external examinations to ensure consistency".[35]

There is no recognition that removing the so-called status distinction between standards-based and traditional examination approaches might mean using unsuitable assessment methodologies for some domains and for some purposes—that there may be a trade-off between equity and efficiency. It is, in any case, unclear what is intended. Does it mean that *All Must Have Prizes*,[36] to quote the title of a recent book, ie that no student should fail and there should be few if any distinctions between achievement levels? It seems to me that any status distinction there may be originates in the world outside education, and there is little that can be done within education to change it, except to ensure that there are rigorous, credible qualifications for a wide range of school-based learning and not just for academic subjects.

I would note that some very high status professions depend on a very high level of knowledge which can be tested on a standards-based approach. I would want to be very confident that the surgeon taking a scalpel to my person knows exactly where *all* the arteries and nerves in the relevant area are located—not just 99 percent of them. Similarly, I want to be assured that the pilots who fly me know how to land and not just how to take off. But these highly developed vocational skills, which can be tested by a pass/fail, standards-based approach, are built on a very substantial foundation of knowledge and thinking ability which is often best tested by traditional methods.

Non-state Examinations

There is also some uncertainty about whether private examinations would be registered on the framework with credits towards the NCEA. The ministry project manager is reported as saying that there was no reason why the International Baccalaureate (IB)

[34] Smith, N. The Hon. (1999), Address to the New Zealand Post Primary Teachers' Association Annual Conference, Wellington, 21 September.
[35] Smith (1999).
[36] Phillips, M. (1996), *All must have prizes*, London: Little, Brown and Company.

shouldn't be recognised: "If a student meets a valid standard and it's a valid standard why shouldn't they [sic] get recognised".[37] This seems to imply that the IB will have to employ standards-based assessment against standards deemed valid by New Zealand authorities. Elsewhere it is stated that the IB and other "registered" examinations that meet the "quality standards defined for the NQF and the NCEA could also give a student credit towards their National Certificate".[38] However, there has been as yet no White Paper on the NQF, so Achievement 2001 is being developed out of the context of the NQF as a whole. We therefore don't know what the "quality standards" are to be. In any case, it is questionable whether the IBO and other overseas examining bodies would change their prescriptions and assessment methodologies to fit in with New Zealand requirements. Nor is it apparent that the award of NCEA credits would in any way enhance an internationally respected qualification such as the IB.

It would seem that New Zealand schools will not have the option of using private examination bodies that do not meet ministry criteria for registration on the NQF and that are not recognised by the award of credits towards the NCEA. All students—no exceptions seem to be envisaged—are to work towards the NCEA.

Outcomes

An Elastic Concept

The Green Paper on the NQF said that: ". . . acceptable outcome statements could include 'purpose statements', 'objectives', 'process objectives', 'standards' or 'learning outcomes".[39] This suggests that the term "outcome" has no clear boundary. There are various descriptions of Achievement Standards. An early pamphlet simply says that they "will specify what students have to achieve in each subject to earn credit towards the NCEA".[40] The Qualifications Development Group (QDG) states that the government has resolved that they are to be "transparent outcomes that students need to meet (that is, based on agreed standards of achievement)".[41] Thus the concept of outcomes continues to be elastic.

The draft Achievement Standards should be the subject of separate analysis. But a brief perusal indicates that at least some are not standards at all and might best be described "general aims". A few level 1 assessment criteria taken at random are:

Health, Food and Nutrition

- Students will be able to:
  - demonstrate a satisfactory knowledge of adolescent nutritional requirements. (credit)
  - demonstrate a thorough knowledge of adolescent nutritional requirements. (merit)

---

[37] Cassie, F. (1999). Overseas exams may get the nod. *NZ Education Review*, June 18.
[38] QDG, 15 June 1999.
[39] Creech (1997), 21.
[40] Ministry of Education and NZQA (1999). *Achievement 2001 – Qualifications for 16 to 19 year olds*. November.
[41] QDG Update of 15 June 1999.

History

* Some ideas(s) or point(s) of view are accurately identified. (credit)
* Most idea(s) or point(s) of view are accurately identified. (merit)
* [The learner] accurately identifies the historical facts that are consistent with the evidence selected from the sources. (excellence)

Science (complete a practical investigation)

* Scientific plan chosen to collect data would be suitable for a trial and after discussion with the teacher (of a trial) the plan becomes appropriate for the purpose. (credit)
* After some trialling or research the scientific plan should be appropriate for the purpose of the investigation. (merit)
* Merit PLUS the plan will include some discussion on the strength and/or limitations of the methodology chosen. (excellence)

Clearly these are not "standards": words such as "satisfactory", "thorough", "some", "most", and "suitable" are wide open to interpretation. The actual standards, to the extent they exist on paper and not in teachers' heads, will not be "transparent" but will be within assessment guides and other material. However, these examples show something of the attempt by the expert groups, explored by Terry Locke in his examination of the draft English Achievement Standards, to dispense with qualitative judgments and to differentiate between the grades in technicist measurement terms—in these cases by the movements from "data . . . suitable for a trial" to "some trialling", from "satisfactory knowledge" to "thorough knowledge", and from "Some idea(s)" to "Most idea(s)". Again, it is hard to see how they can form an adequate basis for the award of merit and excellence. There is no clear attempt to ask markers to discriminate between candidates in terms of *how well* they have performed in a given task—which is, of course, one of the main limitations of the competency-based approach of Unit Standards.

Can the Concept Apply to All Learning?

The concept of learning outcomes is central to Achievement 2001, as it is for the NQF, and indeed for the Curriculum Framework. The issue was discussed at length in the context of the NQF by Smithers[42] and Hall,[43] but the concerns raised by them do not appear to have been acknowledged, let alone addressed.

Smithers challenged the basic assumption, which is one on which the NQF rests, that all learning can be expressed appropriately as outcomes. He pointed out that school subjects are different ways of making sense of the world and that not all learning is organised as subjects: academic study, applied education and vocational training have different organising principles. While it might well be important in, for example,

[42] Smithers (1997), 39–40.

[43] Hall, C. (1998), *The National Qualifications Framework Green Paper – What future for the framework?* In I. Livingstone (ed.), *New Zealand Annual Review of Education*, 7, Wellington: New Zealand Council for Educational Research.

occupational training to specify what is required in terms of outcomes, other matters such as relevant experience, that is, training, may well also be important. In the humanities the point might be for students to have shared their insights and it is intrinsic to the experiences that students take away different things from them. In such cases the most authentic method of assessment might be to set common tasks (for example, discuss the meaning of a particular play) and "by comparisons of the responses (rather than against pre-set outcomes) judge what the candidates have made of it".

The Relationship of Achievement Standards to the Curricular Outcomes

The Achievement Standards are to be derived from the outcomes of the curricula developed for the various "essential learning areas" established by the New Zealand Curriculum Framework. The curriculum statements are to "specify clear learning outcomes against which students' achievements can be assessed" and also assessment procedures and examples.[44] It is not clear why we need to retain the curricula from Form 5 upwards if we are to have Achievement Standards which also purport to set the desired outcomes against which students' learning can be assessed.

The new curriculum statements have been assessed in Education Forum submissions and reports. They reveal a generally poor set of curricula ranging in quality from "could do better" to downright appalling.[45] The new curricula do not form a good basis for the construction of "standards" for qualifications. The lack of content in many of the curricular outcomes, their vacuity, and the student-centred approach to outcome determination raise serious doubts about the suitability of the curricula as a basis for summative assessment. It is not clear to me how "curricula" which lack content, activities and process can be readily converted into standards or examination prescriptions.

The Effect of Outcome Specification on Pedagogy

The change from content to outcomes as the basis for curriculum specification will, I assume, be having significant effects not only on *what* is taught but also on *how* it is taught—on pedagogy. The change represents a decisive movement towards "the why" and the "who for" and away from "the what" and "the how", that is, away from the questions of "what content should be taught?" and "by what appropriate methods and procedures shall students acquire it?"[46] The pedagogical effects of this movement—good or bad—will be greatly increased in the senior school by replacing examination prescriptions specifying content by "outcomes"—or, perhaps more accurately, by

---

[44] Ministry of Education (1993), *The New Zealand curriculum framework*, Wellington: Learning Media, 1, 22–23.

[45] See Education Forum submissions and reports: Education Forum, *English in the New Zealand curriculum – A submission on the draft* (1994); *Technology in the New Zealand curriculum – A submission on the draft* (1994); *Social Studies in the New Zealand curriculum – A submission on the draft* (1995); *Social Studies in the New Zealand curriculum – A submission on the revised draft* (1996); *Health and Physical Education – A submission on the draft* (1998); *Art in the New Zealand curriculum – A submission on the draft* (1999); Howson, G (1994), *Mathematics in the New Zealand curriculum – A review*, Auckland: Education Forum, August; Kelly, P (1995) *Science in the New Zealand curriculum – A review*, Auckland: Education Forum, November.

[46] For discussion see Education Forum (1999), *Arts in the New Zealand curriculum – A submission on the draft*, Auckland: Education Forum, September, 49–51.

"general aims"—in the form of Achievement Standards. This seems to me to be an important issue which remains to be addressed.

Clearly there can be positive aspects of an outcomes-based education (OBE) in so far as it seeks to make clear what is to be learnt and the stages in which learning can best be undertaken: indeed it makes the obvious and very important point that learning is expected to take place. But its suitability surely depends on what is to be taught. Where outcomes can be readily defined then OBE would seem eminently suitable—but this assumes content, a concept which is regarded by the government as outmoded.[47] Where the outcome is less easily defined, perhaps even ineffable, then clearly OBE is problematic. How, for example, could one specify as "transparent outcomes", and by grade and by level, the qualities that, it is hoped, education in the arts will foster—qualities that defy precise specification such as:

> . . . sensitivity, delicacy, poise, elegance, imaginative insight that can go off on its own, the abil:'y to respond in the appropriate mode, to paraphrase, to simplify and render open for inspection some of the deeper levels of significance that the surface of a work [of art] conceals—above all, to see or present all the many-sidedness of things in one occasion, presentation or object—all the informed understanding and responsive appreciation that furnish the marks of progress and success in being educated in arts subjects, what Broudy (1971)[48] called the "enlightened cherishing" that comes along with, and is promoted by, an education in the arts.[49]

Size of Outcomes

As noted above, the size of Achievement Standards is to be larger than that for Unit Standards to avoid the problem of atomisation. A Ministry of Education pamphlet of 1 June 1999 says they are to be "broad" and many who were concerned about the application of Unit Standards to "conventional" school subjects will welcome this. But at the same time it is difficult to see how one can have "breadth" and "transparency". Isn't there a tension here? A subsequent ministry pamphlet tends away from breadth and towards transparency when it states that "A reader [of a student's NCEA] will know *exactly* what a student knows and can do within a subject, and will know which subjects and what parts of the subject the student is very good at, or otherwise" (emphasis added).[50] This would seem to imply the incorporation of an enormous amount of detail into each Achievement Standard and the loss or diminution of any sense of overall achievement: it resonates with the sort of claims that were made of the original NQF.

---

[47] The Green Paper entitled *Assessment for success in primary schools*, May 1998, noted that "In New Zealand, the shift from a content-based curriculum to an outcomes-based curriculum indicates [the] change of focus [from one on inputs to one looking directly at student learning outcomes]" (p. 7).

[48] Broudy, H.S. (1971). Enlightened preference and justification. In R.A. Smith (ed.), *Aesthetics and problems of education*. Urbana, Illinois: University of Illinois Press.

[49] Education Forum (1999), 79.

[50] Ministry of Education (1999), *Policy Directions*, issue 1, July.

I would also note that the rejection of all forms of scaling and the substantial amount of internal assessment envisaged must inevitably put considerable pressure on the Achievement Standards to be very specific in the interests of consistency and comparability of assessment and thus the credibility of the qualification. Internal assessment, for example, means that many different assessors in different schools will be employing different assessment tasks, and students will be assessed under different conditions and rules. Of course, moderation will aim to reduce the inconsistencies and achieve reliability, but there will be a trade-off between central direction in the form of increasingly precise standards and local autonomy in terms of decisions about the assessment tasks to be used, and when and in what manner they are to be undertaken.

Setting Outcomes of Equal Difficulty

One of the least contentious results of assessment research is, I understand, that it is very difficult and maybe impossible to set assessment tasks of equal difficulty from year to year. Many factors are involved, including the wording of the tasks, the options allowed and the order in which tasks are listed. And if the tasks themselves vary not just between years but between schools, and if conditions under which they are taken also vary between schools, then the achievement of comparability becomes even more difficult. Inter-year scaling is designed to address this issue and in most circumstances[51] its results are entirely useful: it allows my Bursary result this year to be compared with Joe's results from last year and Jill's results next year. In Achievement 2001 we are to establish between five and eight Achievement Standards per subject per level, and all at the same level are, presumably, to be of equal difficulty. The same, of course, applies to the sets of standards to be established for each grade (credit, merit and excellence) at each level. Scaling to even out differences in marker toughness is also ruled out. So all this will add still further pressure on the Achievement Standards to be very precise indeed.

The Elusive Goal of Precision

In the circumstances it would not be difficult for the "spiral of specification", about which Alison Wolf wrote in the context of the United Kingdom's National Vocational Qualifications (NVQs), becoming a serious problem. It is worth quoting what she said:

> The 'standards' on which NVQs are based are intended to be so precise that they convey exactly what an assessor should look for. ... [T]his goal of precision has proved elusive. . . . [T]his ever-receding goal of total clarity derives not from bad luck or incompetence, but is actually inherent in the methodology adopted. The more serious and rigorous the attempts to specify the domain being assessed, the narrower and narrower the domain itself becomes, without, in fact, becoming fully transparent. The attempt to map out free-standing content and standards leads, again and again, to a never-ending spiral of specification.[52]

---

[51] Circumstances when it might not be useful include a very small student group taking the examination or when there is some significant change in the composition of the student group.

[52] Wolf, A. (1995), *Competence-based assessment*, Buckingham: Open University Press, 55.

The government's aim has been to avoid the difficulties encountered with Unit Standards including "issues of outcome definition" and of the "credibility of internal assessment".[53] I applaud the intention, but I am not convinced that Achievement 2001 manages to avoid those difficulties: indeed they have been perpetuated and transparency and credibility will continue to be elusive. It looks to me as if we will be revisiting the problems that have been already written about in the context of Unit Standards within the NQF.[54]

Costs and Benefits

I have found no detailed assessment of the costs and benefits of the new system. This should be undertaken and the results confirmed or adjusted by trialling before a decision is made on changes to present arrangements. In this case there is no analysis to speak of and there are to be no trials. What we do have are a great many questions and issues which await resolution.

It is hard to identify the benefits. The one unified system employing the best of what we have at present looks suspiciously like a slightly modified Unit Standards system, the deficiencies of which for much senior school work have been identified in several analyses of the NQF and Unit Standards.[55] The fixation with frameworks has driven policy towards a highly restricted system which will not accommodate much of the learning that takes place in the senior school and will meet few of the information needs of the users of school qualifications.

On the costs side, there would appear to be major increases in store, offset to some small extent by reductions. But it is hard to be specific, given so many uncertainties. Clearly the development costs are considerable and there will be costs on all secondary schools in familiarising themselves with the new Achievement Standards and assessment procedures, as well as, of course, with the new curriculum statements. There would appear to be substantial increases in internal assessment at Forms 5 and 7, with commensurate increases in administration and in moderation. However, given that the Sixth Form Certificate is entirely internally assessed at present, it seems likely that externally assessing at least half will reduce teacher assessment workload at that level, though the reduction may be small because the linkage with the previous year's SC results amounts to a simple "moderation" procedure. As regards the external assessments, presumably the Achievement Standards will replace existing marking schedules or require new marking schedules to be developed. Presumably there will be three or four external examinations per subject (for each of the externally assessed Achievement Standards) and not just one.

If the new system provides qualifications which are not widely accepted nationally and internationally, there will be very major costs. For example, some schools will decide that Achievement Standards are unacceptable and will face the costs of developing alternative qualifications systems or signing up to overseas ones such as the IB—provided, of course, they are allowed to do this; this may require a change in policy which

[53] QDG Update of 15 June 1999.
[54] See, for example, Irwin, M.D.R., Elley, W., and Hall, C. (1995); Elley (1996); Irwin (1997); Smithers (1997).
[55] See Elley (1996) and also Smithers (1997), which lists such material, e.g. on p. 22.

presently requires *all* learning to be towards the NCEA. Schools offering only the NCEA may face a substantial fall in the numbers of overseas students, and consequent loss of revenue from that source, if their parents lack confidence that the NCEA will be readily recognised by New Zealand and overseas tertiary institutions. Tertiary institutions and employers may face the costs of developing and applying their own cognitive assessments to provide the information presently provided by SC, Sixth Form Certificate and Bursary.

The costs on New Zealand students who find that their school qualifications are not readily understood and accepted are incalculable. Whatever new system is introduced, there will be the costs of discontinuity if it is not possible to readily compare the new qualifications with the present SC, Sixth Form Certificate and Bursary. Finally, there will be the costs of further dislocation as problems within the new system arise and lead to significant modification or its replacement.

Conclusions on Achievement 2001

I would summarise my conclusion on Achievement 2001 as follows:

(i) it is not supported by any detailed analysis of:

* the government's aims as regards school qualifications
* the present difficulties and obstacles to be overcome in achieving those aims.

Nor is there any identification and evaluation of the alternative solutions, taking into account the inevitable tensions and trade-offs. It is being developed at enormous speed without trialling. It is a very high risk venture with a great deal at stake;

(ii) the lessons from our own recent experience of the NQF and from experience elsewhere with forms of standards-based assessment (including achievement- and competence-based) have not been sufficiently heeded;

(iii) the policy is being made "on the hoof" as problems, which were entirely predictable and which should have been identified at the policy design stage, are being encountered. It is unsurprising that Achievement 2001 is riddled with contradictions and unidentified and unresolved tensions. At times it seems as if we are expected to be capable, like the White Queen, of believing "as many as six impossible things before breakfast";[96]

(iv) there is no evidence whatsoever that Achievement 2001 will be a better alternative to what we have at present and that the considerable costs of development and implementation and of the dislocation caused among schools and teachers are outweighed by whatever benefits there may be. In fact it is hard to identify any benefits; and

(v) development and implementation of Achievement 2001 should be put on hold pending a thorough, professional analysis of school qualifications drawing on available research from within New Zealand and overseas. In the meantime present arrangements should continue. The last thing anybody should want is

[96] Carroll, L. (1971). *Alice's adventures in wonderland and Through the looking-glass*. Oxford: Oxford University Press, 177.

a major policy failure in this area which is so important to the life chances of so many of our young people.

In a recent analysis of the curricular reforms, I said:

> . . . my overall judgment is that the curricular reforms of the last ten years have imposed enormous costs with few discernible benefits. . . . Given the apparent lack of thought that has gone into them, this is not surprising. I have little doubt that we will spend much further time and effort over the next decade in unraveling them.[57]

It is disappointing to have to come to a very similar judgment on yet another major educational policy initiative and one which also affects so many people.

Where To From Here?

Having come to these melancholy conclusions, it is incumbent on a critic such as myself to indicate, at least in outline, the lines a further enquiry into school qualifications might profitably follow. I stress that the following is "work in progress". It does not seek to provide precise answers, which would require much more research and well conducted trials.

First, Do No Harm

A great deal is at stake with Achievement 2001, for the life chances of generations of pupils, for teachers whose morale is already suffering from an onslaught of policy changes, for school boards and administrators, and for the wider society. Any significant change in qualifications will have substantial costs. Surely the first and obvious point to make is that we should be very confident that the changes are worthwhile, that they represent a significant improvement on what we have now, and that the costs are clearly outweighed by the benefits. At the very least we should adopt the medical prescription: first do no harm. I would urge a cautious, evolutionary approach with well-researched changes being thoroughly trialled before final decisions are made.

The Purpose of School Qualifications

We should consider very carefully the purpose of school exit qualifications. Sometimes it seems to me from the official papers that uniformity and framework construction are ends in themselves. But this is unhelpful and likely to preempt several important issues.

An obvious but necessary point to make is that qualifications can only certify certain aspects in the cognitive domain—certainly not everything about a student and not even everything he or she knows or can do. Other methods of reporting are required for the professional judgments and intuitions of teachers about their students' attributes, non-classroom performances, personalities, dispositions and likely areas of success beyond school. This can include all sorts of information, for example about participation in

[57] Irwin (1999a).

school leadership, sporting and cultural activities, and can be supplemented by portfolios, certificates and such like.

Secondly, in talking about qualifications we are primarily concerned with summative assessment, not diagnostic, formative or evaluative assessment.[55] The value of school exit qualifications lies in what can be done with them—in the doors and pathways they open up at the tertiary level or in employment. This necessarily requires that qualifications discriminate in terms of cognitive abilities and achievement levels.

The importance of distinguishing between students is enormously increased by the sheer numbers of students now sitting SC and staying on to the sixth and seventh forms. Having SC or UB no longer indicates that the holder is one of an elite group at or near the top of the age cohort in terms of academic ability. Employers and tertiary institutions need to distinguish between the tens of thousands who now take the secondary school qualifications. For example, in 1980 there were 8400 UB candidates. This grew to 19,400 in 1990 and to 27,400 in 1998.

Parents also need to be able to monitor secondary schools' performance in terms of the qualifications achieved by their pupils. For this they too will need methods of aggregating the results of individuals' achievement so as to compare schools' results. Of course, parents take other matters into consideration when choosing and monitoring a secondary school, but for most academic achievement is a key aspect.

With these huge increases in the number of qualifications holders, the qualifications will become stratified; even those held out by education authorities to be equivalent will be ordered by users within a hierarchy. This may be a complicated and messy business, but students along with tertiary institutions and employers will quickly learn to "read" the system.[56] Success in maths may, at least for some purposes, be seen as more important than apparently equivalent achievement in social studies. Externally assessed components may be seen as more credible and therefore rated more highly than those that are internally assessed. Users appreciate that underneath a norm-referenced examination result lies the notion of standards. Over time they learn to understand what such results indicate in terms of actual ability levels.

If users of qualifications can't select with confidence on the basis of school qualifications, they will either institute their own examinations or use what they know about the schools the applicants attended—which is likely to be at best a rough and ready guide and to the disadvantage of able students attending less well-favoured schools.

What Is To Be Assessed?

The next question is, surely, "What should be assessed?". I have already queried why we should have two sets of "outcomes"—one in the curricular statements and another in the Achievement Standards—which both purport to set transparent standards against which students can be assessed. We should, I suggest, scrap the curricula for Form 5 and upwards. Whatever is prescribed for the qualifications will drive the curriculum anyway.

---

[55] With more and more pupils staying into the sixth and seventh forms, it can reasonably be argued that assessment in the fifth form is increasingly becoming formative rather than summative.

[56] See Wolf, A. (1996), Education: The Tyranny of Numbers, *Prospect*, December.

How should the qualification requirements be specified? The new curriculum statements and the Achievement Standards purport to be outcome statements (though may be more accurately described as statements of aims), which means that curricula are, in effect, left to schools to determine. But this creates a major problem for national assessment: if there are no common curricula, how can we have common assessment tasks? In any case, our current obsession with outcomes ducks one of the most important questions in education—the choice of content, which, as Peters has remarked, is:

> . . . between a range of activities that are thought to be worth passing on. Science, mathematics, history, art, cooking and carpentry feature on the curriculum, not bingo, bridge and billiards.[60]

Content cannot, as seems to be implied in some of our new curricula, simply be left to student choice. As the Education Forum said in a recent submission:

> . . . explicit reference to, and specification of, suitable content are . . . vital if the moral agency of the school as an educating institution and the professional integrity and disciplinary responsibility of the members of its teaching staff are to have any purpose and point of application to students' learning.[61]

Thus from both philosophical and pragmatic points of view, a return to prescriptions specifying content would seem to be essential.

Which Assessment Methodologies?

Determining the appropriate assessment methodologies seems to me to be the next task. Surely it is quite unhelpful to address this issue before specifying what is to be assessed. And it is quite wrong to specify one "unified" methodology as if everything that needs to be assessed in the senior school is amenable to the same methodology. To put it another way, specifying only one methodology limits what information can be obtained and what can be assessed—starting with assessment methodology is to start at the wrong end. It seems to me that if we are to assess the diverse knowledge and skills which are now taught and learnt in the senior school we will have to accept that there have to be dual (if not triple or quadruple) assessment approaches. And there may well be a trade-off here between manageability and the number of different methodologies that should, ideally, be employed.

If there are status issues to be addressed, I suggest they will be better addressed by ensuring that there are high quality vocational qualifications and alternative "general" papers for those whose aptitudes and aspirations are unlikely to lead them to academic work at the tertiary level. Issues of status cannot just be wished away by prescribing one "unified" form of assessment—this is fanciful and unhelpful.

---

[60] Peters, R.S. (1966), *Ethics and education*, London: George Allen and Unwin (p. 144).
[61] Education Forum (1989), *The arts in the New Zealand curriculum*, Auckland: Education Forum, October, 51.

Where information about relative ability is sought, it seems to me likely that there will be a strong bias towards norm-referenced examinations, which will not always require written responses, employing common assessment tasks, which are externally set and marked, and undertaken under common conditions. The problem of cheating is largely avoided thereby and thus credibility ensured. There is also a substantial benefit for schools in that it avoids the heavy costs of internal assessment and moderation.

Where it is possible and desirable to assess against a set standard it may well also be important to employ external examiners to reduce the possibility of cheating and thus to ensure credibility. As a general rule, the higher the stakes the more important is credibility, and thus the more important assessment using common tasks under common conditions by independent examiners becomes. In so far as summative assessment at Form 5 is decreasingly important as an exit qualification, there is more of a case for internal assessment, and for those who take Bursary this applies for Sixth Form Certificate as well. However, students become better at examinations with practice, and the sooner they start the better.

We need to rescue "scaling" from the list of educational swear words. Of course, the various forms of scaling can be abused, but they can also be very useful and sometimes essential if the aims of assessment are to be achieved. Although officially denied, a form of inter-year scaling continues with SC and for the very good reason that, as noted earlier, it is impossible to set questions of equal difficulty each year. Because of the aggregation of subject scores, inter-subject scaling continues in Bursary as does inter-year and, as necessary, inter-marker scaling. It is high time the forms of scaling were explained and their use defended.

Reporting Results

What is reported and the manner of doing so must reflect the needs of the users of qualifications. What is presently intended is unclear, but it seems that it will involve reporting on five to eight Achievement Standards per subject taken for Forms 5 to 7. A student taking an average of six subjects and staying on to Form 7 is likely to end up with reports on about 100 Achievement Standards. Users will not be able to cope with such numbers and will select, perhaps discounting all but the most recent results and perhaps only looking at some subjects and perhaps only the externally assessed Achievement Standards within them. In other words it seems likely that there will be an excessive amount of reporting on current proposals. What is needed for most subjects are fewer, aggregated scores—not a vast amount of detail, which might well present a confusing picture of students' actual abilities and relative levels of attainment.

As a general rule, users of qualifications will want measures of students' abilities in activities in which they, the users, are most interested. According to Wolf:

> If you are trying to predict whether people will be good at higher levels of academic study, academic tests are therefore appropriate. Conversely, if you are interested in vocational or professional work, or in practical skills, then they are almost certainly *inappropriate*. The evidence . . . strongly confirm[s] the limitations of academic assessment as predictors of later performance. It also showed that assessments which mirror closely the

> activities and environment of a workplace are good predictors of workplace success.[62]

This points to the different information needs of the users of qualifications. Universities, for example, will seek potential to undertake academic work at high levels. They will generally seek indications of relative ability and, especially in selecting entrants to high demand courses such as medicine, need fine distinctions in performance, not simply three broad grades. On the other hand, some employers and those tertiary institutions running more vocational or occupational courses may well be interested in evidence of success in particular activities which may be provided by some form of standards-based assessment.

For academic courses it will generally be the case that overall judgments by experienced, well-socialised markers and based on broad criteria are appropriate. For occupational courses evidence of success in quite precisely defined activities may be needed. One form of assessment will not suffice.

Inappropriately sized criteria will have unfortunate effects on instruction as well as on assessment. As Cedric Croft has noted:

> Standards that are too specific may have an adverse impact on learning; but standards that are too broad probably lead to assessment judgments that are unreliable.[63]

Assessment methodology and the manner in which the material to be assessed is specified are bound to affect instruction and learning. Qualification design should seek where possible to ensure that assessment raises achievement levels and not simply records what levels have been reached.

Separation of Curriculum from Assessment and Qualifications Design

One issue which has not been addressed in any of the papers that I have seen is the administrative arrangements for curriculum and qualifications, though it will presumably be addressed in the White Paper on the NQF. In my view, under whatever arrangements are devised, the tasks of setting curricula and of assessment for the senior school should be separated. The curricula should be set by subject working groups under the direction of a curriculum authority, and assessment cum qualification design should be undertaken by examination bodies. At present we seem to have a very unsatisfactory situation in which curriculum and assessment decisions, for example about which standards should be externally assessed, are jumbled up in one process, with predictably muddled results.

If we effect this separation there would be no prior decisions about which methodology to use and about how much of each subject should be internally and externally assessed. Such decisions would be made by examination bodies based on what methods are best suited to the material to be assessed and the purpose of the assessment. I used the plural—examination *bodies*—advisedly because I do not think we should become locked into one qualification system. Only if there are two or more bodies

[62] Wolf (1995), 129.
[63] Croft (1994), 18.

offering their services can we allow innovation and see which bodies offer the kinds of services schools, parents, students, employers and tertiary institutions actually want. The "one-best" system is a "take-it-or–leave-it" system in which innovation is difficult and which forces students to take qualifications which may not suit their post-school aspirations. It is also one in which political considerations will always tend to trump educational ones.

Can Qualifications be Combined Within a Framework?

There is an increasing range of senior school courses as schools seek to make schooling at the senior levels relevant to a wider range of students in terms of their abilities, attainments and post-school aspirations. These necessarily result in a wider range of qualifications—no longer only in conventional academic subjects—and this raises a complication for framework designers. To what extent can diverse qualifications be sensibly aggregated so that the aggregations, and therefore the students, can be readily compared? Reducing everything to a common currency of credits is a superficially attractive solution. But qualifications are likely to be based on different assessment systems to meet different purposes and to certify different types of ability. And this makes the establishment of a common currency problematic—how, for example, could a standards-based Unit Standard be aggregated in any sensible fashion with the marks from a norm-referenced examination?

Users may not accept the credit rating attached to some qualifications and may even disregard some credits entirely from their calculations. In other words, education authorities cannot expect qualifications users to rate qualifications as they think they should be rated. It is better to simply list qualifications in cases where aggregations would be forced and artificial. To do otherwise is to risk bringing the system into disrepute. The notion that a nice, neat qualifications framework can be constructed in which all qualifications have a predetermined value based on a common "currency" of credits derived from a single "unified" assessment methodology which will be accepted by all is, I suggest, an illusion.

What are needed in the senior schools are qualifications catering for the increasingly diverse abilities, attainment levels and post-school aspirations of senior students. At present the senior school is largely geared to the more academically inclined students. Quality programmes leading to quality qualifications in the applied and occupational areas are urgently required. At present there are only Unit Standards. The ministry clearly accepts the shortcomings of Unit Standards but does not propose any changes to improve them, except the possible introduction of a merit grade. Achievement 2001 sticks to a non-differentiated curriculum framework and offers a non-differentiated school qualifications system. This is an enormous opportunity missed.

Conclusion

What I have tried to do in this paper is to analyse present proposals for school qualifications and to suggest an alternative way of thinking about them. In my view Achievement 2001 is a blind alley. It has been motivated by the search for a political compromise, seeks the elusive "one-best" system for the assessment and certification of

all school-based learning, and has been encouraged by the current fixation with educational frameworks. It sought to offer all things to all people but in fact offers a very restricted system littered with contradictions and unidentified problem areas. Moreover, it ignores substantial research that has been undertaken in New Zealand and elsewhere into forms of standards-based assessment, not least research into New Zealand's very recent experience with the National Qualifications Framework.

Further work is urgently called for, and Achievement 2001 should be put "on hold" until it has been undertaken. An enduring, respected, credible school qualifications system will need to take seriously into account epistemological constraints and the limits of assessment technology. A system built on political compromise and not on sound educational foundations may have short-term attractions but will ultimately fail. While the current arrangements could, in my view, be improved (especially in the non-"academic" areas), they are certainly not so hopeless as to require the drastic revolutionary treatment envisaged by Achievement 2001. A cautious, evolutionary approach which identifies and addresses particular problems within an overall view of school qualifications is much to be preferred.

In developing that overall view, the further work required will need to ask "first-order" questions such as the purpose of school qualifications and the information needs of their users. It should not be burdened by a prior commitment to a single "framework" solution to multidimensional problems. Tensions and trade-offs must be confronted and not blurred. The further work needs to go on to consider what should be assessed and only then to ask what assessment methodologies are suitable given the material to be assessed and the information needs of users. I have suggested that an important issue here is the relationship between curriculum on the one hand and summative assessment and qualifications design on the other; in my view they are best separated. The issue of the extent to which qualifications can be linked within a framework is one of the last things to address not one of the first. The approach of Achievement 2001 is back to front.

In short, a great deal more thinking needs to be done before we can proceed with any confidence to reform or replace what we have at present. I hope that what I have had to say is accepted as a useful contribution to that further work.

CONCLUSION

Dr Richard Watkins

Director and Chief Executive
New Zealand Council for Educational Research

CONCLUSION

In its 1999 annual conference, NZCER sought to address the complex and at times confusing issue of assessment. In choosing this topic, NZCER did so because of the closeness of assessment to widely held perceptions about the value of education; the ability of contemporary students of all ages in critical areas such as literacy and numeracy; the seemingly intrinsic relationship between assessment and educational accountability; and because it is an issue which never seems to vanish entirely from any wider debate about education.

NZCER did not set out, as the range of speakers at the conference and the papers herein confirms, to present a particular point of view which painted some types of assessment as desirable and others as unpalatable. It did seek to present a point of view, however, that good assessment for whatever purpose is imperative, and that we should eschew poor assessment under all circumstances.

The conference presentations and the papers in these proceedings make it clear that problems will rapidly emerge if assessment is limited to only minimalist notions, if it lacks clarity of purpose, if teachers and students regard it as a necessary but unwanted evil, and if it fails to make a significant contribution to furthering the laudable aim of raising educational standards, no matter how these are expressed.

One thing which is abundantly clear is that it is impossible to separate good formative assessment from good teaching and learning. This will not surprise any of us. Moreover, that assessment must be clearly linked to educational objectives, not in the narrow sense of instrumental outputs, but in the sense of those broad outcomes which derive from the national curriculum no matter how imperfectly it is presented.

It is also clear that the nation's teachers need continuous training, guidance and assistance in deriving, applying and interpreting good assessment practices, whether these have formative or summative purposes. Again, this should really be no great surprise.

As the array of assessment tasks becomes ever larger, and the responsibilities of teaching become greater, the professional development of teachers is crucial to their understanding and use of good assessment. Thus to fail to provide suitable continuing training for the nation's teachers in selecting, using and analysing appropriate assessment instruments and their results can only be detrimental to the very students we are all concerned about.

In presenting the 1999 conference and these proceedings, NZCER trusts that it will have fulfilled part of its mission: *NZCER - supporting educators through quality research, resources, and information*. The purpose is to improve the quality of education through research; reduce barriers and expand educational horizons for all New Zealanders; and create educational resources based on sound research.